

# Objects predict fixations better than early saliency

**Wolfgang Einhäuser**

**Merrielle Spain**

**Pietro Perona**

Division of Biology,  
California Institute of Technology,  
Pasadena, CA, USA, &  
Department of Neurophysics,  
Philipps-University Marburg,  
Marburg, Germany



Computation and Neural Systems,  
California Institute of Technology,  
Pasadena, CA, USA



Computation and Neural Systems,  
California Institute of Technology,  
Pasadena, CA, USA



Humans move their eyes while looking at scenes and pictures. Eye movements correlate with shifts in attention and are thought to be a consequence of optimal resource allocation for high-level tasks such as visual recognition. Models of attention, such as “saliency maps,” are often built on the assumption that “early” features (color, contrast, orientation, motion, and so forth) drive attention directly. We explore an alternative hypothesis: Observers attend to “interesting” objects. To test this hypothesis, we measure the eye position of human observers while they inspect photographs of common natural scenes. Our observers perform different tasks: artistic evaluation, analysis of content, and search. Immediately after each presentation, our observers are asked to name objects they saw. Weighted with recall frequency, these objects predict fixations in individual images better than early saliency, irrespective of task. Also, saliency combined with object positions predicts which objects are frequently named. This suggests that early saliency has only an *indirect* effect on attention, acting through recognized objects. Consequently, rather than treating attention as mere preprocessing step for object recognition, models of both need to be integrated.

**Keywords:** attention, eye movements, object recognition, scene recognition

**Citation:** Einhäuser, W., Spain, M., & Perona, P. (2008). Objects predict fixations better than early saliency. *Journal of Vision*, 8(14):18, 1–26, <http://journalofvision.org/8/14/18/>, doi:10.1167/8.14.18.

## Introduction

What guides attention? Although the concept of selective visual attention dates back at least to the late 19th century (James, 1890), the factors driving this selection process are still far from understood. Two distinct questions are of interest: First, what is the role of top-down factors (e.g., task, observer idiosyncrasies) as compared to what can be inferred from the stimulus (bottom-up factors)? Second, what is the role of low-level features—such as contrast, color, orientation, flicker, motion—as compared to higher-level stimulus structure—such as objects or gist? In the present study, we focus on the second question, utilizing eye movements as correlates of the focus attention (cf. Rizzolatti, Riggio, Dascola, & Umiltà, 1987). Specifically, we ask whether fixations are driven directly by “early” (low-level) saliency or through correlations to higher-order scene structure, such as the saliency of recognized objects. In other words, is attention driven by mechanisms

that are earlier than, and independent from, recognition, or is attention part of the recognition process itself?

Most attention models are based on a so-called saliency map (Itti & Koch, 2000; Koch & Ullman, 1985). Filtering the input image with kernels reminiscent of early visual mechanisms generates feature maps at various spatial scales. These are then combined into a single saliency map, which encodes the probability that an image region will be attended. The saliency map is entirely based on early features and was originally designed to explain covert attention on simple stimuli. Surprisingly, however, saliency maps predict, to some extent, fixations also in complex scenes (Parkhurst, Law, & Niebur, 2002; Peters, Iyer, Itti, & Koch, 2005; Privitera & Stark, 2000; Tatler, Baddeley, & Gilchrist, 2005). Some authors hope that, by progressively refining such low-level models, human attention will eventually be modeled perfectly. In this view, attention operates independently of object recognition and may be thought of as preceding and guiding object recall. This view has recently been challenged.

Even if features of the saliency map, such as luminance contrast, are good *correlates* of fixation probability (Krieger, Rentschler, Hauske, Schill, & Zetzsche, 2000; Mannan, Ruddock, & Wooding, 1996, 1997; Reinagel & Zador, 1999), a number of authors have argued that they do not necessarily drive attention *causally* (Carmi & Itti, 2006; Einhäuser & König, 2003; Tatler, 2007) but contingent on higher-order statistics (Einhäuser, Rutishauser, et al., 2006). Rhesus monkeys preferentially fixate image regions with semantic content as compared to meaningless (noise) regions with the same low-order statistics (Kayser, Nielsen, & Logothetis, 2006), and it has been suggested that objects, such as faces, may drive attention in a direct fashion (Cerf, Harel, Einhäuser, & Koch, 2008; Hershler & Hochstein, 2005, 2006; but see vanRullen, 2006). Along similar lines, the “perceptual experience” rather than the stimulus per se pre-dominantly influences eye movement behavior when viewing art that has ambiguous experiences (Tatler, Wade, & Kaulkard, 2007). Therefore, even in the absence of an explicitly formulated task, eye movements are to a large extent influenced by higher order scene properties, and scene interpretation.

The fact that the specifics of the task influence eye motions had been noticed as early as Buswell (1935). In his seminal study, Yarbus (1967) used a variety of tasks, including abstract interpretations, such as the judgment of social status. In these cases, the task clearly dominates the fixation patterns, as it does in complex activities of daily living (Land & Hayhoe, 2001). Recent studies suggest that during visual search, early saliency has only a minor or no impact on fixation patterns (Einhäuser, Rutishauser, & Koch, 2008; Henderson, Brockmole, Castelano, & Mack, 2006; Underwood, Foulsham, van Loon, Humphreys, & Bloyce, 2006), and the effect of a stimulus feature on fixation depends on its relation to the search target (Pomplun, 2006). Models that modulate low-level channels attempt a mechanistic explanation for such top-down regulation (Navalpakkam & Itti, 2007; Rao, Zelinsky, Hayhoe, & Ballard, 2002; Tsotsos et al., 1995). This highlights that “bottom-up” and “low-level” are to be carefully distinguished.

In addition to task and stimulus features, search in natural scenes is influenced by prior knowledge on the typical spatial location of the search target, as well as by contextual information. Modulating saliency map models with such priors improves their fixation prediction (Torralba, Oliva, Castelano, & Henderson, 2006). Even beyond search, such spatial priors may influence fixation behavior. The “central bias” of observers, the tendency for observers to fixate preferentially close to the center of photographs of natural scenes, might in part reflect the expectation of interesting objects in this region (for a detailed account on the factors possibly contributing central biases, see Tatler, 2007). In this view, spatial priors are believed to be a bottom-up function of scene statistics that is learnt from experience and applied in a task-dependent (top-down) manner.

Based on James’ (1890) original notion that attention “implies withdrawal from some things in order to deal effectively with others,” it is generally believed that attention’s main function is the allocation of processing resources to accomplish complex tasks such as visual recognition. In this view (“attentional bottleneck”; Nakayama, 1990), attention precedes recognition in the processing pipeline. The precise relation of attention and recognition, however, is largely unresolved. On the one hand, it has been argued that rapidly recognizing the “gist” of a scene does not require *spatial* attention (Li, VanRullen, Koch, & Perona, 2002; Rousselet, Fabre-Thorpe, & Thorpe, 2002). On the other hand, a variety of phenomena point to an involvement of attention in recognition and/or recall,<sup>1</sup> such as inattention blindness (Neisser & Becklen, 1975; Simons, 2000), change blindness (Rensink, O’Regan, & Clark, 1997), repetition blindness (Kanwisher, 1987), and the “attentional blink” (Raymond, Shapiro, & Arnell, 1992; for natural scenes, see Einhäuser, Koch, & Makeig, 2007; Evans & Treisman, 2005).

The extent to which overt attention, or attention associated with shifts of gaze, is needed to recall an item has been studied extensively. Friedman (1979) demonstrated that unexpected items are fixated longer and recalled better; similarly, Nelson and Loftus (1980) show that—in particular for brief presentations—change detection requires close fixations. Hollingworth and Henderson (2002) find an advantage for detecting changes in items fixated previously and a correlation between the time spent fixating an item before the change and change detection. Consistent with these results, Tatler, Gilchrist, and Land (2005) show that the information of an object’s position is accumulated over fixations but do not find a similar effect for object identity. In addition to this better memory for fixated items, it has been argued that changed items are also fixated earlier after the change (Parker, 1978). Henderson, Williams, Castelano, and Falk (2003) challenged these findings based on experiments using objects embedded in a more complex background, and instead they find change detection—and thus the guidance of attention—restricted to a small region around the current fixation. Although the precise details of the relation between fixation and memorization seem dependent on experimental paradigms, all these data suggest that there is some relation between the allocation of overt attention and the ability to recall certain properties of an item.

Attention-free feed-forward systems—no matter whether designed to optimize performance or to model physiology—perform particularly well on categorical recognition tasks when the scene is pre-segmented into patches containing a single object category (Einhäuser, Hipp, Eggert, Körner, & König, 2005; Fei-Fei, Fergus, & Perona, 2004; Fergus, Perona, & Zisserman, 2003; LeCun, Bottou, Bengio, & Haffner, 1998; Mel, 1997; Riesenhuber & Poggio, 2002; Serre, Wolf, Bileschi, Riesenhuber, & Poggio, 2007; Wallis & Rolls, 1997). However, in the real

world, objects rarely occur in isolation, but rather in the midst of clutter, and may cover as little as 0.1% of the image area (Rutishauser, Walther, Koch, & Perona, 2004). Then attention might be a necessary preprocessing step for recognition (Dickinson, Christensen, Tsotsos, & Olofsson, 1997), for learning new objects, and may speed up recognition (Rutishauser et al., 2004). In summary, while psychophysical evidence suggests that spatial attention is not needed for recognizing isolated objects or the gist of isolated scenes, attention most likely supports recognition in spatially and temporally cluttered settings. The precise interaction of attention and recognition in natural conditions is thus of interest for human and machine vision.

In sum, the literature suggests that saliency maps based on early visual features have some, albeit limited, power in predicting eye movements and attention in natural complex scenes. This limit is likely intrinsic; that is, higher-level visual properties of the scene will have to be considered in order to see a significant predictive improvement. While it is clear that objects such as faces have the power to draw attention, we are still far from a quantitative model that predicts eye movements from the configuration and visual properties of objects in a scene. In this study, we attempt a step in this direction: We explore which object properties drive attention. Furthermore, we test the hypothesis that the most “meaningful” object in an image attracts attention and, once one takes this effect away, raw saliency maps have little predictive power.

To explore the relation between attention and recognition in a natural setting, we use semantically rich natural photographs (Shore, Tillman, & Schmidt-Wulffen, 2004). To ensure observers’ alertness, while preserving natural viewing behavior, all observers are asked to evaluate the aesthetic value of each picture. To investigate the effects of visual search on fixation statistics, half of the observers in addition search for a verbally defined target object. In both conditions, after object search and aesthetic evaluation, we ask observers to characterize scenes with “keywords” in order to measure which objects were seen and remembered as significant. For both conditions, we assess the mutual relation between three quantities: the locations our observers fixate, the locations of objects they recall, and the locations of highest saliency according to the Itti and Koch (2000) model. This allows us to compare how well each of five different quantities (raw saliency, object saliency, an optimal combination of both measures, the mutual prediction of different observers, and general spatial biases) predict fixations.

## Methods

### Stimuli

The stimuli were 93 photographs from the artist S. Shore’s collection “Uncommon Places” (Shore et al.,

2004; Figure 1). The images were collected as a “visual diary” and come across as casual snapshots of everyday scenes. The images were presented on a 20-inch CRT monitor, located in a dark room at 80 cm from the observer, and thus subtended  $29 \times 22$  degrees of visual angle. The artist provided digitized high-resolution images. To fit the resolution and the aspect ratio of our presentation screen ( $1024 \times 768$  pixels), images were down-sampled and cropped (minimally).

### Experimental conditions

We tested two experimental conditions, referred to as “what” and “where.” In both conditions, we instructed our observers to imagine that they are a “judge for an art competition” and to rate, on a scale from 1 to 5, “how interesting” each image was. Asking our observers to rate the images insured that they would observe them carefully. We did not use the ratings in our analysis.

In both conditions, observers were asked to provide “some (up to five) keywords” to describe the scene. To avoid confounding the eye-tracking data, the keywords were typed after the stimulus disappeared and after the observers provided the aesthetic rating. In the “where” condition, observers additionally searched for an object (the target), which was specified in writing on the screen before image presentation. Observers were asked to decide as quickly as possible whether the target object was present in the scene. Target objects were chosen to make search difficult; targets were either present but not obvious (this was established independently in a non-eye-tracking Internet-based “what” condition), or not present in the image, but plausible for the scene and frequently named in other images (Table 1).

In the “what” condition, each image was displayed for 3 s. In the “where” condition, the image disappeared as soon as observers responded about object presence by pressing a key. Following the disappearance of the image, the observer rated its “interestingness” from 1 to 5, following which the observer typed up to five keywords (Figure 2A).

### Observers

Eight volunteers (6 male, 2 female; mean age: 23) from the Caltech community participated for pay, four in each condition. All participants were native English speakers, had normal or corrected-to-normal vision, and normal color vision as assessed by Ishihara plates. None of the participants had any formal art training. All were naive to the experiment’s purpose and had not previously seen the stimuli. All procedures conformed to National and Institutional Guidelines for experiments with human subjects and to the Declaration of Helsinki.





Figure 1. Stimuli. Ninety-three photographs of Stephen Shore's collection "Uncommon Places" were used as stimuli (reprinted with permission of the artist).



Image	Recalled objects
1	woman (7); street (4); sidewalk (3); fence (3); bag (2); clothes (1); purse (1); plant (1); bush (1); building (1); glasses (1); heel (1)
2	cantaloupe (7); pancakes (7); water (3); butter (2); food (2); knife (2); milk (2); table (2); drink (1); juice (1); plate (1); syrup (1)
3	floor (7); plant (7); trash can (4); wall (4); pot (2); smoke alarm (2)
4	chair (6); TV (5); desk (4); bed (2); pitcher (2); wall (2); closet (1); door (1); glass (1); heater (1); lamp (1); mirror (1); screen (1); shelf (1)
5	painting (7); lake (5); wallpaper (5); trees (4); mountain (3); drum (1); eagle (1); headdress (1); Indians (1); rocks (1); shoreline (1); frame (1)
6	cars (7); lake (5); cloud (3); parking_lot (3); sky (3); tree (3); storm (2); sand (1)
7	car (7); building (6); road (6); shop (3); ad (1); paper box (1); Pepsi (1); sign_1h_parking (1); sign_no_parking (1); sign_spruce (1); stoplight (1); tree (1)
8	road (7); cars (4); traffic light (3); building (2); water (2); curb (1); headlight (1); hydrant (1); light pole (1); shutters (1); sign (1); tree (1)
9	church (8); car (5); woman (5); street (3); antenna (2); building (2); beetle (1); cross (1); sky (1); steps (1); sunday_only_sign (1); wires (1)
10	house (5); road (5); car (3); buildings (2); chimney (2); smokestack (2); factory (1); hill (1); lot (1); roof (1); sign (1); telephone booth (1); weeds (1)

Table 1. Recalled objects for first 10 images of Figure 1. Number in parenthesis provides recall frequency of the object; objects are sorted by recall frequency. A table with all keywords is available as supplementary material at <http://www.staff.uni-marburg.de/~einhaeus/download/ObjectRecall.csv>.

## Recording eye position

Throughout the experiment, a non-invasive infrared Eyelink-1000 (SR Research, Osgoode, ON, Canada) system monitored eye position at a 1000-Hz sampling rate. Our analysis used only data recorded during stimulus presentation (Figure 2B). The chin and forehead rests of the system stabilized observers' heads. The calibration of the eye tracker's gain was validated after each 10 trials and recalibrated when necessary. Linear drift of the eye tracker was controlled for before each trial onset and corrected when needed. The average validation error in a 13-point validation procedure was  $0.56^\circ \pm 0.10^\circ$  (mean  $\pm$  SD over subjects). This error is on the order of the saliency maps' resolution (1/16th of the image resolution, i.e.,  $0.5^\circ/\text{bin}$ ) and smaller than the typical object size, which we coarsely estimate by the square root of the number of pixels covered by an object, yielding 223 pixels or  $6.3^\circ$  on average.

Thresholds to detect saccades were set to a velocity of  $35^\circ/\text{s}$  and an acceleration of  $9500^\circ/\text{s}^2$  as recommended by the manufacturer for the Eyelink-1000 device. There was no minimum duration for a fixation set, but 99.4% of the 7318 fixations lasted longer than 50 ms and 97.6% longer than 100 ms (median: 251 ms; mean: 311 ms). The location of a fixation was defined as the mean eye position during this fixation. The maximum horizontal distance covered by the eye during a fixation was below  $0.5^\circ$  in 79.0% of cases, below  $1^\circ$  for 98.0% of fixations, for the vertical direction these values were 80.5% and 96.7%, respectively (mean:  $0.37^\circ$  and  $0.38^\circ$ ; median:  $0.32^\circ$  and  $0.31^\circ$ ). The standard deviation of a fixated location was on

average  $0.08^\circ$  both in horizontal and in vertical direction. The typical variation of fixated location during a fixation is thus small compared to the absolute location accuracy of the eye tracker, the resolution of saliency maps, and the typical size of objects.

Presentation of stimuli, recording of eye position, and analysis were implemented in Matlab (Mathworks, Natick, MA) using its psychophysics and eyelink toolbox extensions (Brainard, 1997; Cornelissen, Peters, & Palmer, 2002; Pelli, 1997, <http://psychtoolbox.org>).

## Object annotation

For consistency of the main analysis, the authors marked the outlines of the objects named by the observers (Figure 2C). For analysis, we excluded terms describing the full image, objects not present, words other than concrete nouns, and repetitions (but counted them in the object naming order). Obvious synonyms were treated as the same object. The image annotation was blind with respect to the fixations, that is, only the keywords and the images were used during synonym determination and object outlining.

To obtain an independent set of labels, we asked—for a subset of images—an additional observer to outline “all objects” in a given image. Since “all objects” is an ill-defined stopping criterion, we motivated this observer to label as many objects as possible by making payment proportional to the number of labeled objects (5 cents/object) plus a bonus for objects that occur in multiple instances in an image (+1 cent/instance).

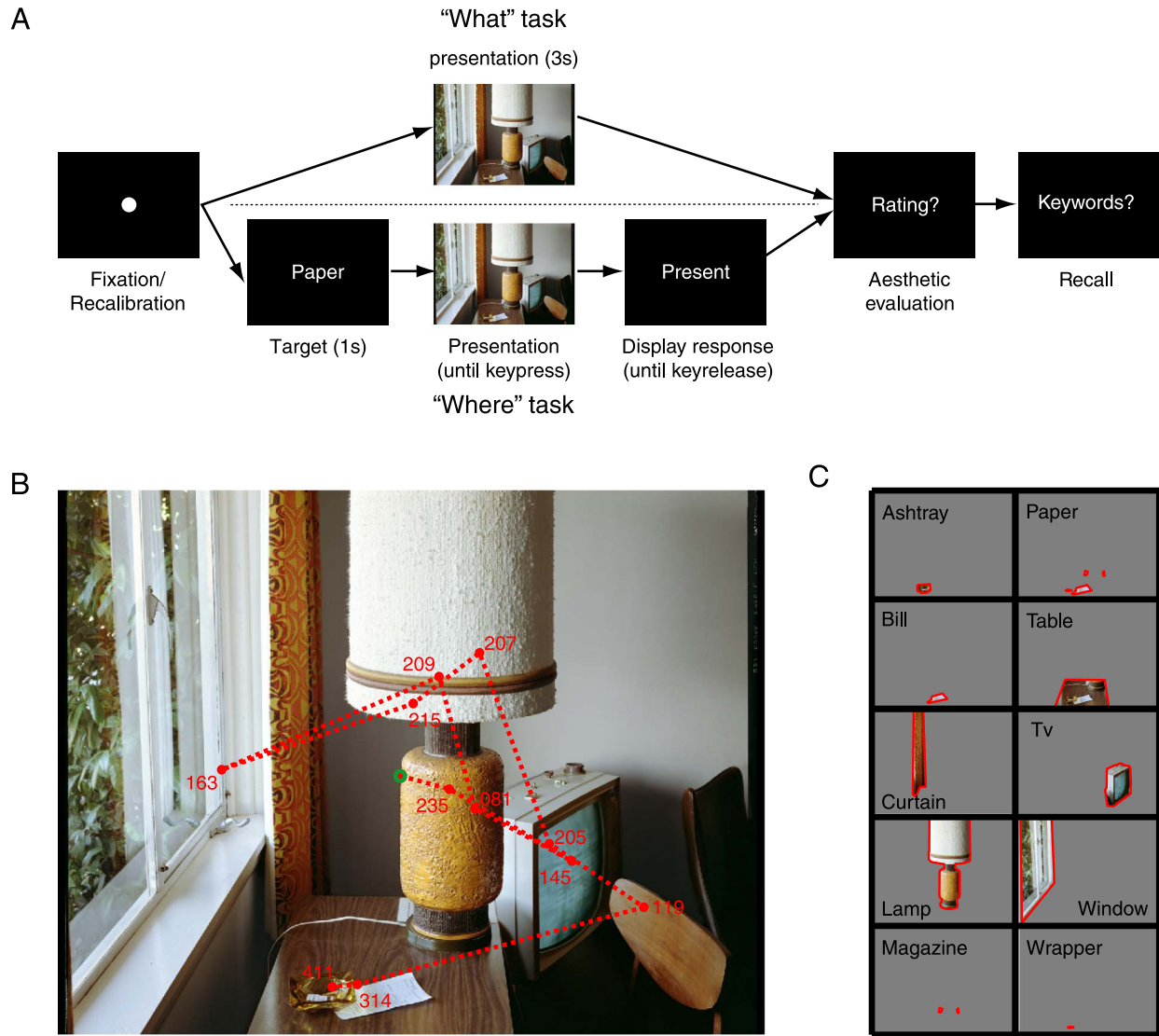


Figure 2. Paradigm and examples for recall Paradigm outline. In the “what” condition (top), observers see the image for 3 s and are prompted for a rating and then for keywords. In the “where” condition (bottom), observers terminate presentation by deciding on the presence or absence of an item (search target) presented verbally at trial onset. (B) Example images superimposed with fixations of a single observer (MW). Numbers at fixations denote fixation duration in milliseconds. (C) Outlines of all named objects for images of panel B.

## Object maps, fixation maps, early saliency, and object saliency

Our definition of “saliency maps” follows the model of Itti and Koch (2000), with the authors’ original parameters and their implementation, which we obtained from <http://ilab.usc.edu>. The computed saliency map has a lower resolution than the original image and is scaled by a factor of 16 (linear) to obtain the map  $S_i(x, y)$  for each pixel of image  $i$ .<sup>2</sup> Analogously, we define an “object map”  $O_i(x, y)$ : For each observer, we count the number of objects overlapping with pixel  $(x, y)$  in image  $i$ . Then we sum these maps of all observers to obtain a single map for each image  $i$ , and finally normalize the map divisively to maximum 1, yielding  $O_i(x, y)$ . Note that in this default

definition  $O_i$  depends on the frequency of recall: An object recalled by all observers is weighted 8 times stronger than an object named once. The term “object map” refers to this (“observer-weighted”) definition, unless stated otherwise. In addition, we consider object maps that are not weighted by the number of observers but count the number of objects overlapping with a given pixel irrespective of the number of observers recalling the object (“unweighted object map”). Both maps are normalized divisively to maximum 1 to ease comparison without affecting the relative ranking of pixels in each map. To test the consistency of observers’ fixations, we define a “fixation map”: we assign each fixation to the nearest pixel and label the respective pixel as fixated. Due to the high resolution of the image, overlap between two fixations on the pixel level can be neglected

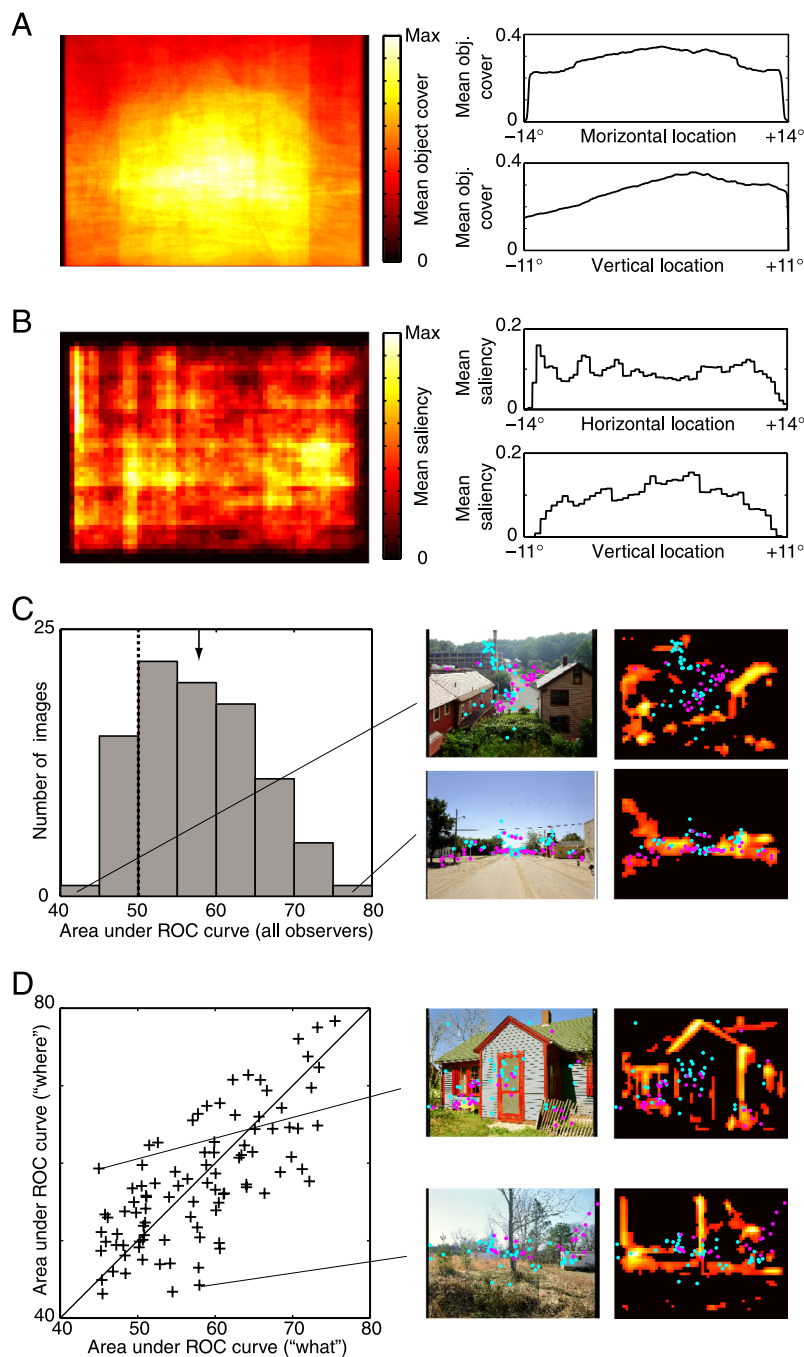


Figure 3. Fixation prediction by saliency maps. (A) Average object map and its mean along the cardinal axis. Note that each of the 93 maps is normalized to the same integral before adding. Other normalization schemes, however, yield qualitatively the same result: a clear maximum to the horizontal center and below the vertical midline. (B) Average saliency map and its mean along the two cardinal axes. Note that there is no pronounced central bias to saliency. (C) Area under ROC curve for saliency map's prediction of fixated locations pooled over all observers in each image, histogram over images. For 77/93 images, prediction is better than chance, arrow indicates mean area under the curve. Example images and saliency maps for images with best (bottom) and worst (top) fixation prediction. Color code for all saliency maps as provided in panel A. (D) Area under the curve separated for "what" and "where" observers. Each data point corresponds to one image; for points above the diagonal saliency map's prediction is better in "where" (45 images), below the diagonal in the "what" task (48 images). Example images and saliency maps for two data points marked by thin lines.

and we obtain a binary map, with entry “1” for fixated and “0” for non-fixated pixels. This map is then smoothed with a  $1^\circ$  Gaussian kernel to obtain the fixation map. Figures 3 and 4 depict examples of object maps and saliency maps and Figure B1 an example of a fixation map.

We define the “total object saliency” of an object as the sum of saliency map values over the object’s footprint divided by the sum across the whole image. Since this measure scales with the area of the object, but the area cannot be factored out easily due to the sparseness of saliency maps, we consider an additional measure. We define “maximum object saliency” as the maximum saliency map value inside the object’s outline. As the features of the saliency map are computed early in the visual hierarchy, we will refer to the saliency map values at a given location as “early saliency.”

## Signal detection analysis

### Predicting fixations

We compute how well each of the aforementioned maps predicts fixations by using a method proposed by Tatler, Baddeley, et al. (2005). Given an image, the respective map is computed scaled up to the image resolution where needed. Again, each pixel is either labeled with 1 (fixated) or 0 (non-fixated). We then computed the fraction of fixated pixels, where the map had values above a threshold (hits), and the fraction of non-fixated pixels where the saliency map had value greater than the same threshold (false alarms). We plotted hits versus false alarms while varying the threshold from zero to one (minimum and maximum values of the saliency map) obtaining an ROC curve (receiver operating characteristic). The area under the ROC curve (AUC) quantifies the quality of saliency’s prediction of fixation. Although other measures of fixation predictions have been proposed in the context of saliency maps (such as, e.g., “normalized scan-path saliency” of Peters et al., 2005), our signal detection measures have the advantage that monotonic scaling of maps does not affect their results. This is especially valuable when the predictions of different maps need to be compared.

### Predicting recall

The prediction of object recall cannot be tested directly since objects that are not recalled by any observer remain unknown. Instead, we tested how well fixated locations discriminate between objects recalled by one observer (idiosyncratic objects) from objects recalled by multiple observers. We label the objects by the number of observers recalling them,  $l(o) = “1”$  for idiosyncratic objects,  $l(o) = “2+”$  for objects recalled by two observers or more,  $l(o) = “3+”, \dots, l(o) = “8”$ . The fraction of fixations inside each object pooled over all observers is used as measure  $f(o)$ . The fraction of objects with label 2 above a threshold  $t$  ( $f(o) \geq t$ ,  $l(o) = 2$ ) are

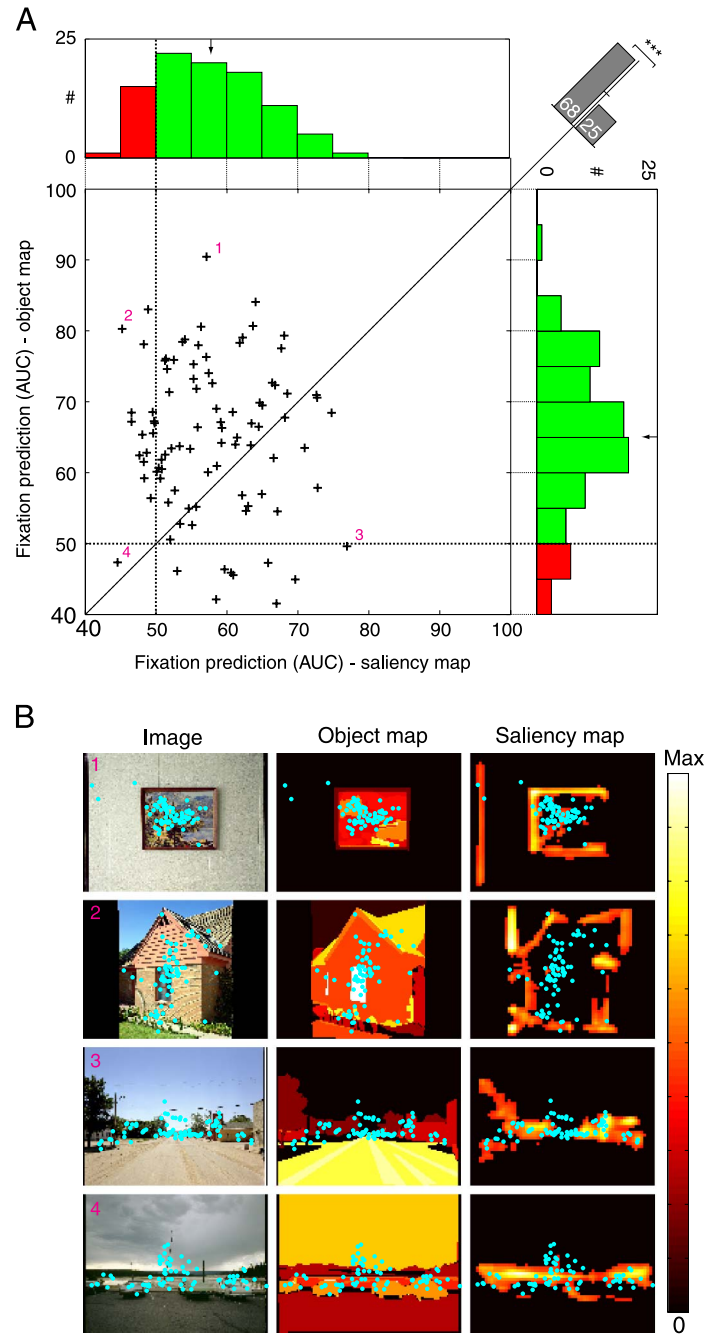


Figure 4. Object maps predict fixations. (A) Area under the curve (AUCs) for fixations predicted by saliency maps (x-axis) and object maps (y-axis). Each data point corresponds to one image. Distribution of either AUC depicted as marginals (same axes as scatter plot). For points above the diagonal object map’s prediction is better than the saliency map’s (68 images), below the diagonal the opposite is the case (25 images). Magenta numbers identify examples in panel B. (B) Examples of images, in which fixations are predicted best by the object map and reasonable by the saliency map (top), well by the object map despite bad prediction by saliency map (2nd from top), best prediction by saliency map, despite bad prediction by object map (2nd from bottom) and bad prediction by both (bottom). Left: image; middle: object map; right: saliency map. Fixations of all observers in cyan.



counted as “hits,” the fraction of objects with label 1 above the threshold ( $f(o) \geq t$ ,  $l(o) = 1$ ) as false alarms. By varying  $t$ , we obtain an ROC, quantified by the area under the curve (AUC). This AUC quantified how well fixations discriminate between objects recalled once from objects recalled twice or more. We performed the same analysis for objects recalled  $n$  times or more as compared to recalled once (objects recalled more than once but less than  $n$  times were excluded for this analysis). With the same analysis, we tested—in addition to  $f(o)$ —the recall prediction of the time of fixations on the objects, object area, length of the object’s boundary, object saliency, and linear combinations of these measures.

## Random reassignment baseline

Fixation patterns are not only driven by image specifics but also subject to spatial biases that are independent of the specific image (Tatler, 2007). The central bias, the tendency to look straight ahead in head-fixed settings, is well known (see also Appendix A; Figures A1C and A1D). This bias would predict that objects placed in the center of the photographs would be fixated more often in our experiments, regardless of their intrinsic importance. As a prototypical example, it is well established that the relation between luminance contrast and fixations is partly due to such a double spatial bias (Einhäuser & König, 2003; Mannan et al., 1997; Tatler, 2007; Tatler, Baddeley, et al., 2005).

We follow two strategies to assess the effect of these spatial biases: First, we directly measure the spatial biases of the feature under investigation (see Figures 3A and 3B). Second, we define a *random reassignment baseline* to measure how much of the prediction by a certain map can be explained by its image-independent spatial biases: We reassign randomly the map under investigation (object map/saliency map/fixation map) of one image to another image. At the same time, we keep the property to be predicted (fixations/object recall) with the image they were actually obtained from. On these surrogate data, analysis is performed identically to the actual data. Any effects arising from general biases in the feature are also reflected in this baseline, while any effects beyond the baseline are image-specific.

## Results

In this section we will show that

1. Fixations are predicted better by objects than by early saliency.
2. When object locations are given, saliency contributes little extra information to fixation prediction.

3. Object saliency predicts how frequently an object is recalled.

Hence, the dependence of saliency and fixations is “explained away” by the dependencies between saliency and objects and between objects and fixations.

## Image properties, central bias

In previous studies, fixation predictions could often be partly attributed to a double central bias (see Tatler, 2007): Human observers tend to look straight ahead and images taken by human photographers tend to be centered on salient objects. We verified the photographer’s bias in our sample, images of Stephen Shore, considering all 981 objects that were labeled by at least one of the 8 observers (cf. Appendix C). We define the center of an object as the center of mass of all its pixels. Half of the objects have their center in a circle of  $6.1^\circ$  radius around the image center compared to the image width of  $29^\circ$ . That is, 50% of object centers fall within a central circle whose size constitutes 18.8% of the image area. This central bias occurs primarily in the horizontal direction: Half of the objects are closer than  $\pm 2.9^\circ$  to the vertical midline of the image, a rectangle that corresponds to 20.2% of the image area. A similar result is observed when replacing the object’s center by its entire “footprint,” represented in the object maps: The average over all object maps exhibits its maximum horizontally in the image center, while the vertical peak is below the midline (Figure 3A). Hence, there is a spatial bias on object location. Note that the spatial bias is enhanced by the fact that—at least in artistic western photography—objects are rarely cutoff at image boundaries (and if so, only pixels within the image would be considered), and that objects that span large parts of the scene necessarily have their center of mass close to the image center. Since the present study does not aim at understanding the origin of photographer’s bias, however, it has to be considered as a property of our stimulus material, regardless of origin and in line with other stimulus sets used in the literature.

In contrast, the averaged saliency map does not exhibit a pronounced peak toward the center. Instead, the saliency distribution is rather uniform if one ignores the boundaries where saliency is zero for technical reasons (Figure 3B). We conclude that, for our stimuli, saliency—on average—has no pronounced bias toward the center of the image.

## Early saliency and fixations

### *Early saliency predicts fixations only poorly*

In this section, we assess how well saliency maps predict fixations. Basic fixation statistics, such as duration and spatial distribution exhibit the expected dependence on task (Appendix A): In the “where” task, fixations last

shorter and are more widely spread. In some images, saliency is an excellent predictor of fixated locations (for details, see [Methods](#)), while in other images prediction is poor; the right panel of [Figure 3C](#) shows the extreme examples of good and bad predictions. When pooling over all observers' fixations, the saliency map model's prediction is better than chance (50%) in 77/93 images. The mean area under the ROC curve is  $57.8\% \pm 7.6\%$ , significantly different from chance ( $p = 5 \times 10^{-16}$ ,  $t$ -test). To understand the meaning of this number, we compute the random assignment baseline as lower bound and the inter-observer prediction as upper bound.

To account for possible effects of spatial bias, we compute a “random reassignment” baseline (for details, see [Methods](#)), as has been suggested earlier (Einhäuser & König, 2003; Mannan et al., 1997; Tatler, Baddeley, et al., 2005). We superimpose fixations from one randomly chosen image on the saliency map of a different image. An effect due to biases unrelated to that particular image would still show up in this baseline. AUCs for this setting reach  $52.9\% \pm 5.7\%$ . Although this number is significantly larger than chance ( $p = 3 \times 10^{-6}$ ), it is significantly exceeded by saliency prediction's value of  $57.8\% \pm 7.6\%$  ( $p = 2 \times 10^{-6}$ ,  $t$ -test). Hence, the prediction of fixations by saliency is not a consequence of a general spatial bias alone.

As upper bound, we measure the fixation consistency of distinct observers. The fixations of one observer are predicted by a map generated from the fixation of all others with an AUC of on average 88.9% (for details and task breakdown, see [Appendix B](#)). This number is far above the 57.8% obtained for saliency, which suggests that fixation prediction by saliency maps, albeit better than random, is far from optimal.

### Task independence of saliency map predictions

Several recent studies (Henderson et al., 2006; Underwood et al., 2006) suggest that saliency maps do not predict fixation in search tasks. As discussed above, we find some predictive power, although certainly not much of it. Across our set of images, we do not find the prediction to be generally better for “what” than for “where,” although the differences in prediction performance can be substantial for individual images ([Figure 3D](#)). Therefore, across our set of object-rich images, there is no evidence for saliency maps *generally* predicting fixations either better or worse in search tasks than in free-viewing for recall.

## Objects and fixations

We now explore an alternative hypothesis: Observers fixate objects rather than salient regions in the image. If true, saliency maps might predict fixations indirectly,

if objects tend to be more salient than background, rather than because fixations depend directly on early saliency.

### Predicting fixations with object maps

To test how well objects predict fixations, we define an “object map” in analogy to the “saliency map” for each image (for details, see [Methods](#)). The object map predicts fixated locations above chance in 83 images, with a mean AUC of  $65.1\% \pm 10.6\%$ , which significantly exceeds chance ( $p = 5 \times 10^{-24}$ ,  $t$ -test; [Figure 4A](#)). This is not fully explained by general spatial biases, as it exceeds the baseline of random reassignment of object maps and fixations ( $59.8\% \pm 10.7\%$ ) significantly ( $p = 0.001$ ,  $t$ -test). When comparing the predictions of object map and saliency map for individual images, the object map outperforms the saliency map in 68 images, while the opposite is the case in only 25 images ([Figure 4A](#)). Note that the ROCs, which are computed individually on each image, are independent of any absolute value of the maps (or of any strictly monotonic mapping to them), which makes this direct comparison possible. A sign-test shows that this fraction (68:25) is highly significant, even when ignoring the absolute size of the effect ( $p = 9 \times 10^{-6}$ ). The default object map is weighted by the number of observers recalling an object. If instead the object map is based on the number of objects overlapping with a given pixel, the mean AUC drops to  $61.9\% \pm 10.5\%$ . This is significantly below the value for weighted maps ( $p = 0.04$ ) but still significantly above the saliency maps' fixation prediction ( $p = 0.003$ ). Image-by-image comparison shows that even the unweighted map outperforms raw saliency in 57/93 images, again a significant fraction (57:36,  $p = 0.04$ , sign test). Consequently, in most images, knowing the objects is more predictive of fixations than only knowing early saliency, even if the recall frequency of objects is unknown.

### If objects are known, early saliency contributes little to fixation prediction

Object naming frequency predicts fixated locations in images. On average, this prediction is better than that of early saliency (for extreme examples, see [Figure 4B](#)). Does saliency contribute any information besides what objects tell us already? And vice versa? As first quantification, we ask how much a linear combination of object maps and saliency can improve fixation prediction. Each pixel in the image  $i$  has a value for the object map  $O_i(x, y)$  and the saliency map  $S_i(x, y)$ . To account for their correlation (as saliency maps predict object recall), we treat  $O_i$  and  $S_i$  as dimensions of a plane, for which each original pixel forms a data point. Note that both maps are by definition normalized to the same dynamic range (0 to 1). For these data, we perform principal component analysis (PCA) and project the values on the principal

axis. By reassigning the spatial coordinates, we obtain a single map  $P_i(x, y)$ . This map is the linear combination of object and saliency maps that accounts for most of the variance. Performing the signal detection analysis on this map yields a performance of  $65.0\% \pm 11.6\%$ , which is indistinguishable from the performance of the object map alone ( $p = 0.995$ ,  $t$ -test) but significantly better than early saliency alone ( $p = 10^{-6}$ ). The optimal linear combination of object and saliency maps is provided by Fisher's linear discriminant analysis (LDA). In analogy to  $P_i(x, y)$ , we compute a map  $L_i(x, y)$  by projecting on the most discriminative dimension (with respect to the labels fixated/non-fixated for each pixel  $x, y$ ). By construction, the prediction of this map for each image is better than the best of the individual maps. Nevertheless, the average AUC over all images of  $69.5\% \pm 8.2\%$  is only 4.5% (percentage points) larger than the prediction by the object map alone. Hence, the optimal linear combination of early saliency and object map is only slightly better than the object map alone. Conversely, the optimal linear combination exceeds the AUC of saliency alone by 11.7%. This shows that early saliency does not add substantially to fixation prediction once recalled objects are known, while object maps are informative even when raw saliency is already known. Note that we did *not* separate training and test set, that is, potentially overfit the data. Hence, 69.5% presents only an upper bound to the predictive power of the combined map on novel data. This is a strong indication that knowing saliency provides at most little extra information, once the objects are known.

### Predicting fixations with object saliency

Next we perform an alternative analysis to test whether saliency provides extra information on fixation probabilities beyond that already provided by object outlines. We combine object and saliency maps by computing “object saliency maps” in 4 different versions: We flood the object footprint with the maximum saliency map value inside the object (i.e., with the object's “maximum object saliency”) or with the total saliency map value inside the object (its “total object saliency”). In one condition (“observer weighted”), we weigh the object with number of observers recalling the objects as in the case of object maps. In the other condition (“unweighted”), the recall frequency is ignored. For “observer-weighted” maps, the fixation prediction is indistinguishable than for object maps alone (65.1%, see above): maximum object saliency results in  $65.1\% \pm 10.9\%$  AUC ( $p = 0.98$ ,  $t$ -test) and total object saliency in  $62.8\% \pm 12.0\%$  ( $p = 0.18$ ). In case of unweighted maps, the numbers drop to  $63.3\% \pm 11.4\%$  and  $62.3\% \pm 11.7\%$ , respectively. These values fall between the results for weighted and unweighted object maps but are indistinguishable from either (comparison to weighted OM:  $p = 0.26$  and  $p = 0.09$ ; comparison to unweighted:  $p = 0.40$  and  $p = 0.82$ ). This strengthens the

result that—once the objects are known—saliency contributes little extra information to fixation prediction.

### Predicting recall with fixations

Next we consider how well fixations predict object recall. For all analysis, we split the objects into 8 categories, depending on the number of observers who name the object. For details on recall statistics and their relation to recall order, the reader is referred to [Appendix C](#). First, we first pool fixations over all observers. The fraction of fixations that fall inside an object's boundary correlates with naming frequency ( $r = 0.44$ ,  $p = 7 \times 10^{-49}$ ; [Figure 5A](#)) as does the relative time spent inside the object ( $r = 0.43$ ,  $p = 3 \times 10^{-45}$ ). Frequently fixated objects are recalled more often. Using signal detection analysis, we compute how well the fraction of fixations inside an object discriminates objects recalled exactly once from objects recalled  $n$  times or more (for details, see [Methods](#)). The fraction of fixations inside the object predicts whether an object is named twice or more (“2+”) compared to exactly once with an AUC of 70.3%. Objects named once are discriminated from objects named 8 times with an AUC of 90.4% ([Figure 5B](#)). This prediction is slightly better in the where task (67.2%, 76.3%, and 81.1% for 1 vs. 2+, 1 vs. 3+, and 1 vs. 4, respectively) than in the what task (67.2%, 72.3%, and 76.1%), but in general fixations predict recall well.

Since fixations are collected from the same individuals as recalled objects, one could argue that the relation between object maps and fixations just reflects the fact that fixated objects are recalled better. We test how well object maps obtained from a subset of observers predict the fixations of a different observer. As a baseline, we first predict fixations of each individual (instead of pooled fixations) by the full object map collected from all 8 observers and average for each image over the 8 resulting AUC values. As expected, the mean AUC over observers is very close to that of the pooled fixations (mean over images:  $64.9\% \pm 10.8\%$ ,  $p = 0.94$ ,  $t$ -test). More importantly, excluding the map of the observer, whose fixations are predicted, does not impair the result significantly (mean  $64.5\% \pm 10.8\%$ ,  $p = 0.77$ ). This shows that the predictive effect of object maps is not contingent on including a particular observer's fixations. To verify this further, we asked a single observer to label “all objects” in the 10 images of best object map prediction. This observer was given unlimited time and paid based on the amount of labeled objects. (Note that overlap of different objects prevents even the map of a single observer from being binary and from simply converging to uniformity.) As expected, the prediction of this individual's object map is worse than the 8-observer object map for all (10/10) images tested. However, the prediction of the individual's object map is still better than chance in 10/10 and better than that of the saliency map in 9/10 images. This shows



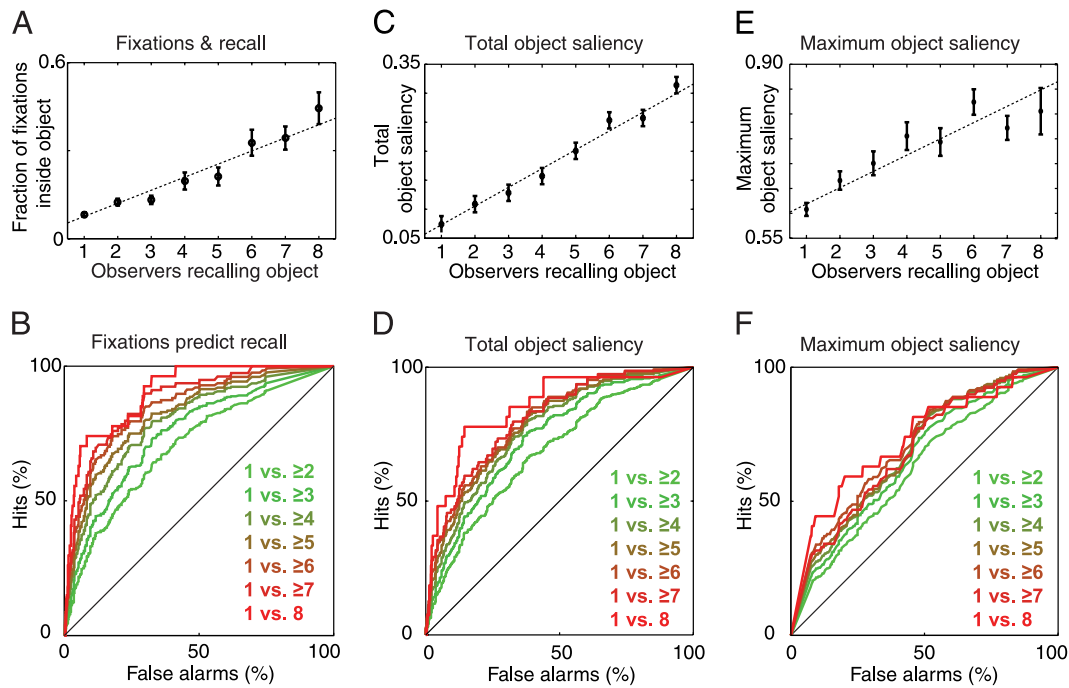


Figure 5. Predicting recall. (A) Fraction of fixations inside object versus recall frequency. Mean  $\pm$ SEM across objects for display; fit treats each of the 981 objects as individual data point. (B) ROC curves separating objects recalled once from objects recalled twice or more, three times or more, etc., using fraction of fixations on object. (C) Object saliency plotted versus its recall frequency. Mean  $\pm$ SEM across objects for display; fit treats each of the 981 objects as individual data point. (D) ROC curves in analogy to panel B, using object saliency. (E) As panel C for *maximum* object saliency. (F) As panel D for maximum object saliency.

that the predictive effect of the object maps is not contingent on the map resulting from the same observer or limits on labeling time. In summary, although fixations and recall are coupled, this effect is not observer specific. Rather than recalling an object because of having fixated it, “interesting” objects (i.e., those frequently recalled) are fixated frequently, even when fixations and recall come from different observers.

## Saliency and object recall

So far we have shown that saliency maps predict fixations to a limited extent, and “interesting” objects (i.e., those recalled by many) are preferentially fixated. Next we aim at completing the argument that saliency maps predict interesting objects and thus predict fixations *indirectly*. The missing part has recently been suggested (Elazary & Itti, 2008) but needs to be demonstrated for our data and conditions: How well do saliency maps predict object recall, how do their predictions compare with the predictions of other object properties, and do their predictions extend beyond fixation alone?

### Object saliency predicts recall frequency

We assign each object a relative “total object saliency,” defined as the sum of saliency map values over the object divided by the sum across the whole image. Across all objects and observers, object saliency is highly significantly correlated to recall frequency ( $r = 0.38$ ,  $p = 2 \times 10^{-34}$ ; Figure 5C). Does this imply that object saliency predicts recall frequency also well on an object-by-object basis? As previously, we perform signal detection analysis, testing how well objects named once can be discriminated from objects recalled more often. Based on object saliency, objects named by all observers are distinguishable from those named once with an AUC of 85.0%, and even objects named twice or more are distinguishable from those named once by an AUC of 68.2% (Figure 5D). Note that this is only slightly worse than the prediction by the fraction of fixations inside the object (Figure 5B). Hence, object saliency of an object predicts its recall frequency nearly as good as the fraction of fixations on the object. For the “what” task, we find AUCs of 68.9%, 72.0%, and 76.6% for distinguishing an object that is named by exactly one observer from those named by two observers or more, named by three observers or more, and named by all four observers. The

results for the “where” task are only slightly different and the differences do not have a consistent sign (AUC: 66.8%, 72.2%, and 74.6%). This shows that object saliency’s prediction of recall is not task dependent.

Since total object saliency scales with object size, we also consider “maximum object saliency,” the maximum saliency map value inside an object. Although the correlation between maximum object saliency and recall frequency is lower than for total saliency ( $r = 0.25$ ; Figure 5E), it is still larger than all the other measures besides object area and is still highly significantly different from 0 ( $p = 1.2 \times 10^{-15}$ ). Similarly, prediction by maximum object saliency reaches AUCs from 62.3% (1 vs. “2+”) to 72.9% (1 vs. 8) and is thus lower than for total object saliency but still substantially above chance (Figure 5F).

Besides object saliency, several other measures suggest themselves for predicting object recall. For *object location*, there is a highly significant correlation between the mean horizontal distance of an object to the image center and its recall frequency ( $r = -0.20$ ;  $p = 2 \times 10^{-10}$ ), whereas the vertical distance exhibits no significant correlation ( $r = -0.04$ ,  $p = 0.19$ ). Besides proximity to the center, *object size* seems an intuitive factor for recall. Recall frequency is significantly correlated to the area covered by the object ( $r = 0.32$ ,  $p = 2 \times 10^{-24}$ ) and the length of the object’s boundary ( $r = 0.22$ ,  $p = 8 \times 10^{-12}$ ). Although this indicates that observers preferentially recall large, central objects, the correlation between total object saliency and recall frequency (Figure 5C) exceeds all other measures tested. These object measures are correlated with object saliency measures and thus partly redundant in predicting object recall. In Appendix D, we show that total object saliency alone is a better predictor of recall than the combination of all other measures and combining other measures with saliency only marginally improves prediction. Consequently, knowing object saliency allows a good prediction as to how often an object is recalled.

To what extent are object saliency and fixations redundant in predicting naming frequency? Figure 6A depicts the relation of total object saliency and fraction of fixations inside an object. Combining the measures along the principal axis of all data slightly improves discriminating rarely named objects from others but does not yield improvements when discrimination is already good (Figure 6B). Similarly, maximum object saliency and fixations on an object are related in predicting recall (Figure 6C), but adding the knowledge about fixations does not add much (Figure 6D). Interestingly, combining maximum object saliency with object area does not reach the levels of total object area, which argues against the effect of total object saliency being a mere consequence of its correlation to object area. This implies that frequently named objects are distinguished from rarely named objects on the basis of maximum or total object saliency and knowing the fixations provides little extra information.

Interestingly, for objects that are recalled by only one observer (“idiosyncratic objects”), the fraction of fixations inside the object is consistently low. Only 25% of such objects have a fixation fraction above 8.9% or below 0.9%. In contrast, for objects recalled by all observers this range extends from 15.8% to 70.4%; that is, the mid half of data covers more than half of the possible range of values. In general, objects recalled by many observers are much more spread out with respect to the fraction of fixations than objects recalled by few (Figure 6E). A similar tendency is observed for total object saliency (Figure 6F). It is tempting to speculate that objects recalled by many observers do not require a fixation to be recalled, while a fixation is necessary to recall objects that are recalled by few. In this view, objects recalled frequently would be named because they are diagnostic for a scene or consistent with its general context, while lesser named objects are primarily recalled as a consequence of fixation. If this hypothesis holds true, the probability to fixate an infrequently recalled object should be larger for the observers recalling it (“recalling observer (s)”) than for the other observers (“non-recalling” observers). This difference should be less pronounced for more frequently named objects. Of the 457 idiosyncratic objects, the recalling observer fixated 188 (41.1%). This compares to 33.6% of non-recalling observers fixating the same objects. Hence, for idiosyncratic objects, recalling observers are about 22.5% more likely to fixate the recalled object than non-recalling observers. The symmetric situation is constituted by the 52 objects that have 7 recalling and 1 non-recalling observers. Here 78.3% of the recalling observers fixated the object, compared to 73.1% for non-recalling observers. Hence, for frequently recalled objects, recalling observers are only 7.1% more likely to fixate the object than non-recalling observer. Similar patterns arise if the fraction of fixations inside the object is considered instead of binary fixated/non-fixated split: For idiosyncratic objects, the fractions are 10.5% for recallers compared to 7.9% for non-recallers, an increase of 32.9%. For objects recalled by 7 observers, the increase is merely 7.9% (34.9% compared to 32.3%). The increase in fixation fraction from non-recallers to recallers and is anti-correlated with the overall number of observers recalling the object (1,...7) with  $r = -0.85$  ( $p = 0.02$ ). Consequently and consistent with the hypothesis, the relative benefit of fixation for recall reduces with increasing number of recalling observers.

### **Saliency predicts a scene’s most characteristic object**

As described before, all object saliency measures are correlated to other object properties (e.g., object area). To obtain an estimate of the effectiveness of saliency in identifying relevant objects in a scene, a more direct measure therefore is to ask whether saliency can predict which of the objects is recalled most frequently (“most

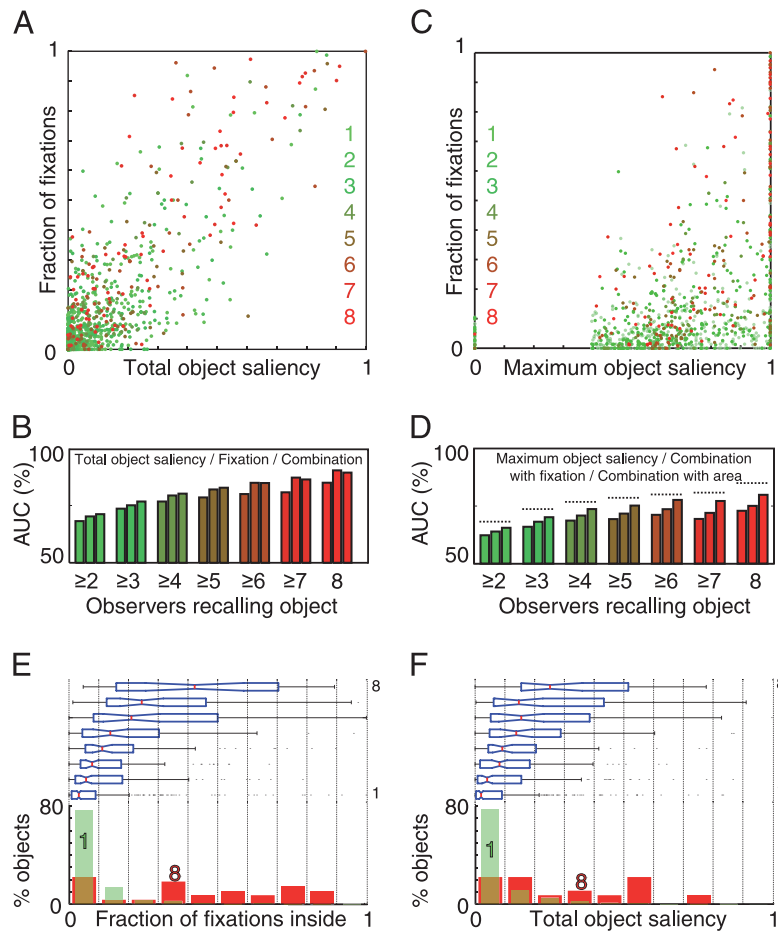


Figure 6. Recall prediction by combination of object properties. (A) Fraction of fixations on an object plotted against its total object saliency, color denotes recall frequency. (B) AUC for recall prediction on the basis of total object saliency alone (*left bars*), fixations alone (*middle bars*), or the combination of both (*right bars*). (C) As panel A for maximum object saliency. Note that maximum saliency frequently takes the extreme values 0 and 1. (D) AUC for prediction by maximum object saliency (*left*), maximum object saliency combined with fixations (*middle*), and maximum object saliency combined with object area (*right*). Dotted lines replicate the results for total object saliency from panel B. (E) *Bottom*: normalized histogram fraction of fixations inside object boundary for objects recalled by 1 (green) or all (red) observers; *top*: boxplots fixations inside object for objects recalled by all (top) to 1 (bottom) observers. Both panels share the same horizontal axis. Note the increase of percentiles from idiosyncratic to diagnostic objects. (F) As panel E for total object saliency.

characteristic object[s]”; see [Appendix C](#)). In 34/93 individual images, the object with highest total object saliency is among those named most frequently (i.e.,  $\text{argmax}_j N(j, i)$  for image  $i$ ). The most frequently named object is unique in 77/93 images (in all but one image, no more than two objects share the highest naming count). In 28/77 of these images, the object most frequently named has the highest total object saliency ([Table 2](#)). For comparison, we measure the probability of obtaining this result through random selection. By performing 10000 simulations<sup>3</sup> of this drawing process, we estimate the expected numbers to be 11.0/93 and 7.7/77, more than 3-fold below the actual values. The maxima obtained across these 10000 simulations are 25/93 and 19/77, respectively. This indicates that the probability to obtain the actual numbers of 34/93 and 28/77 at random is far below 1/10000 (i.e.,  $p \ll 10^{-5}$ ). Hence, total object

saliency predicts the most frequently named object significantly better than a purely random selection that assumes a uniform probability over the image and object properties. Since total object saliency factors in object area, we also tested maximum object saliency. The object with highest maximum object saliency is among the most frequently selected in 35/93 and 26/77 images ( $P_{93}(X \geq 35) \ll 10^{-5}$ ;  $P_{77}(X \geq 26) \ll 10^{-5}$ ). Remarkably, 9/26 (13/35) of these objects were not selected by total object saliency ([Table 2](#)).

How does the object saliency measure compare to other measures in predicting the most frequently named object? The largest object is among the most frequently named in 22/93 (16/77) images, which is still significantly better than chance (simulations:  $P_{93}(X \geq 22) = 0.001$ ;  $P_{77}(X \geq 16) = 0.004$ ) but more than 50% exceeded by the 34/93 and 28/77 of saliency. Similarly, proximity to the image center is not as predictive (23/93, 18/77;  $P_{93}(X \geq 23) =$



Image number	Most named	Named by	Highest total object saliency	Highest maximum object saliency	Hit by saliency peak (baseline)	Largest area	Closest to center	Longest boundary
5	Painting	7	X	X	X (0.50)	X	X	
12	House	7	X					
15	Parking lot	7	X					
16	Car	8	X	X		X	X	
19	Trailer	7	X				X	
20	House	7	X	X	X (0.63)	X	X	X
22	House	7	X	X	X (0.30)		X	
24	Man	8	X			X		X
25	House	7	X	X	X (0.26)	X		
26	House	8	X	X		X	X	
27	Chair	8	X	X	X (0.36)	X	X	
29	House	6	X					
40	TV	8	X	X	X (0.25)		X	
41	House	7	X	X	X (0.42)	X	X	X
48	Building	8	X	X	X (0.14)			
49	Companion	5	X	X	X (0.56)	X		
53	Building	6	X	X	X (0.73)	X		
54	House	7	X	X	X (0.67)	X	X	
59	House	8	X	X	X (0.49)	X	X	
60	Woman	7	X	X	X (0.31)			
63	Car	7	X			X		X
65	Car	8	X					
71	Pool	8	X			X		
74	Tree	6	X			X		X
77	Man	6	X	X	X (0.31)			
83	Field	4	X					
84	Shed	8	X				X	
85	Bed	8	X	X	X (0.20)		X	
4	Chair	6		X				
9	Church	8		X	X (0.30)			
14	Café	6		X				
58	Flag	5		X	X (0.07)			
64	Bush	4		X	X (0.31)			
68	Ford sign	6		X	X (0.04)			
69	Pool	8		X	X (0.68)			
81	Team	7		X				
91	Light bulb	8		X	X (0.12)			
31	Road	8				X		
18	Puzzle	7					X	
46	Lamp	8					X	X
73	Mailbox	8					X	
80	Car	7					X	
90	House	7					X	
46	Parking lot	7						X
Sum			28	26	22	16	18	7
Union						27		

Table 2. Out of the 77 images, which have a unique characteristic object (2nd column), this object has the highest total object saliency in 28 images (4th column), the highest maximum object saliency in 15 (5th column), is the largest in 16 (7th column), the closest to the center in 18 (8th column), and the one with largest boundary in 7 (9th column). The maximum of the saliency map falls on the most frequently recalled object in 22 images (6th column), even if the fraction of image covered by this object may be as small as 4% (number in 6th column).

0.0006,  $P_{77}(X \geq 18) = 0.0006$ ) as saliency. Choosing the object with the longest boundary is indistinguishable from random selection (13/93, 7/77;  $P_{93}(X \geq 13) = 0.30$ ,  $P_{77}(X \geq 7) = 0.65$ ). This shows that although other object properties, such as object size and central biases contribute to object selection, these are exceeded by both total and maximum object saliency.

The most frequently named object is among the set of the largest, most center-proximal, and longest boundary (which can be 1, 2, or 3 distinct objects) in 36/93 (27/77) of the cases, which is comparable (for all images) or even worse (for the 77 with unique most characteristic object) than object saliency (Table 2). In turn, only for 10/59 (7/49) images for which the most frequently named object is not the most salient, any of the other measures predicts this object. This means that only in few images, the other measures can provide information not already contained in object saliency. Consequently, object saliency predicts the most frequently named object better than any other tested measure or any combination of them. More importantly, other measures do not add much, once object saliency is known. In summary, object saliency best predicts which object is most frequently recalled (most characteristic) in each image.

### Maximum saliency falls on frequently named objects

A complementary way of analyzing how well saliency predicts named objects is to ask whether the maximum of the saliency map is located within an object that is recalled, and if so, if these objects include the most frequently recalled. Note that this is different from the maximum object saliency analysis before, as the maximum is now determined over all pixels of the image and there is a possibility that the maximum is not covered by an object. The baseline for this analysis is the probability that the peak of the saliency map covers the object at random, which equals the object area divided by the total image area.<sup>4</sup> The maximum of the saliency map falls on a named object in 78/93 images (83.4%) compared to the mean over all images for the baseline value (mean object coverage) of  $77.0\% \pm 18.7\%$ . To assess significance, we compare the mean of the baseline values (one continuous value per image) to the fraction of images in which the maximum is located within the object boundary (one value for the set) and find them to be significantly different at  $p = 6.4 \times 10^{-4}$  (*t*-test). The most frequently named object encloses the maximum of the saliency map in 29/93 images (31.2%), which is again significantly larger than the baseline ( $22.6\% \pm 19.4\%$ ) of area covered by the most frequently recalled object(s) ( $p = 4.6 \times 10^{-5}$ ). Restricting analysis to the 77 images with a unique most characteristic object, the maximum is in this object in 22/77 images (28.6%) compared to the  $20.6\% \pm 17.8\%$  of area covered by these objects on average ( $p = 1.9 \times 10^{-4}$ ). Table 2 (6th column) provides a list of these objects with the respective baseline values. In summary,

these data show that saliency maps, even without any further knowledge of object content, can be used to pick an image region containing a relevant object better than chance. This reinforces the interpretation of saliency maps as measures of (possibly pre-attentive) scene content.

## Discussion

The present study reconciles two seemingly conflicting views of attention: On the one hand, theoretical models use early features (Itti & Koch, 2000; Koch & Ullman, 1985) and presuppose saliency computation in early visual areas (Li, 2002). On the other hand, there is mounting physiological evidence that saliency is computed later in the visual hierarchy: Frontal areas, such as the frontal eye fields (FEF; Thompson & Bichot, 2005), are known to *represent* saliency. Furthermore, recent microstimulation experiments (Armstrong, Fitzgerald, & Moore, 2006) suggest a direct link from FEF to saliency representation in the visual area V4, which is a prime physiological candidate for saliency *computation* (Bichot, Rossi, & Desimone, 2005; Mazer & Gallant, 2003; Ogawa & Komatsu, 2006). In the light of our results, these views are not conflicting (Figure 7). Raw (“early”) saliency is computed in early visual areas (V1/V2), but by itself has only a small impact on attention guidance (57.8% AUC, see fixation prediction results, black pathway in Figure 7). Instead, early saliency combined with other object properties models the probability of an object being recalled. The location of characteristic objects is then a better predictor of attention (65.1% AUC) than early saliency alone. Furthermore, adding early saliency information to characteristic object location contributes very little predictive power. As shown by the LDA analysis, 69.5% (cyan) is an upper bound on the best linear combination of object footprints and early saliency, when training and test sets were identical, and nonetheless does not substantially exceed the 65.1% prediction of object recall alone. In this view, early saliency does not drive attention directly (black pathway in Figure 7) but through its correlation with object properties (red pathway). In other words, the prediction of objects by saliency in combination with the prediction of fixations by objects explains away the prediction of fixations by saliency. Based on the aforementioned physiological evidence, we may speculate that areas high in the ventral stream, such as V4 or IT, serve as integration site of object recognition and early saliency. Regardless of cortical site, our data indicate that the computation of attention-driving saliency may be distributed but has a component late in visual processing that is relevant for natural scene perception.

The present data and their suggestion that saliency drives attention indirectly through predicting interesting objects reconciles earlier findings: Saliency map features

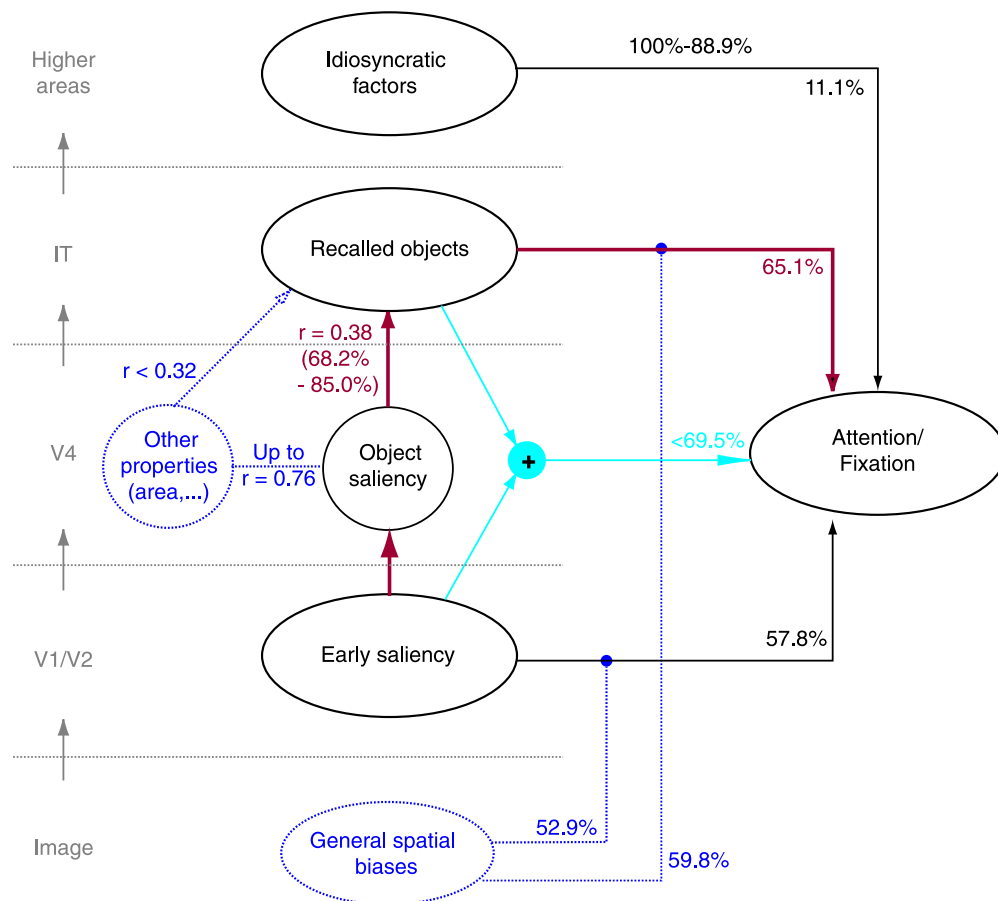


Figure 7. Overview of results. *Right*: Although early saliency predicts fixations to some extent (57.8% average AUC), this prediction is mostly explained through correlations to object recall, as depicted by the red pathway: Object saliency is the integral of raw saliency within the object's boundary, which is highly correlated to recall frequency; the resulting object map then predicts fixations with 65.1% average AUC, which is only slightly below the 69.5% upper bound for an optimal linear combination of both (cyan). The random reassignment baselines reveal that some of the results are accounted for by general spatial biases, which are not specific to individual images (blue). Idiosyncratic factors include everything not explained by the mutual prediction of different observers (88.9% AUC). *Left*: putative brain areas for computation of the individual steps: Early saliency is based on early visual mechanisms, while object representations follow in higher ventral areas. This is consistent with the prime site of “saliency” computation (or integration) being in V4 or IT.

do not need to drive attention (Carmi & Itti, 2006; Einhäuser & König, 2003; Tatler, 2007) despite saliency's undisputed correlation with fixations during free viewing (Peters et al., 2005). Nevertheless, we do not argue that saliency maps are the final answer as to how interesting objects are selected, or that saliency map features would causally drive object recognition. Further research by targeted manipulations of object properties is needed to analyze which stimulus features drive attention, and how they relate to features that make an object “interesting,” “characteristic” or “diagnostic” for a scene and to different types of recall (recalling tokens, types, scene gist, object positions, etc.). Our data suggests, however, that the allocation of attention is preceded by some pre-attentive understanding of the scene. This is in line with the data of the DeGraef (1998, 2005), showing that the even the earliest guidance of attention and fixations depends on whether or not an object is semantically

plausible for a scene. The minimum requirement for such a decision is a coarse pre-attentive recognition of the scene context (or gist) and some form of pre-attentive figure-ground segmentation. Taken together with our present data, this strongly suggests that attention cannot be understood as mere preprocessing step for recognition, but both need to be handled in a common framework.

In earlier studies of eye movements in natural scenes, prediction by saliency maps could often be partly attributed to generic spatial biases in both fixation and saliency. Human photographers typically center objects in images (cf. Figure 3A) and we prefer to look straight ahead, so this “central bias” can artificially enhance measured fixation prediction (Mannan et al., 1996; Tatler, 2007; Tatler, Baddeley, et al., 2005). Here we find that the influence of such double biases is substantial but does not fully explain the observed relations (Figure 7, blue). Furthermore, the bias itself must be represented in the



brain and have adapted to stimulus statistics. Hence, even a stronger bias than the one observed would not invalidate the conclusions regarding the neural computation of attention guidance.

We do not find task dependence of saliency's predictive power for fixation. Two differences with respect to previous studies are obvious: First, our targets are verbally defined preventing observers from knowing their features in advance; second, the locations of our targets are difficult to predict from context, which plays an important role in search (Torralba et al., 2006). So our results do not necessarily conflict with these studies.

The effect of observer idiosyncrasies (e.g., memories, cognitive preferences, etc.) is low for our stimuli and tasks, as reflected by the high inter-observer consistency of 88.9% AUC in mutual fixation prediction. It is well conceivable that this number, which bounds the possible performance of bottom-up models, may drop substantially for different tasks or stimuli. This, however, would only strengthen the conclusion that higher sites are important for driving attention. We stress that the interaction of top-down and bottom-up is not topic of the present study. Instead, we focus on the bottom-up aspect in evaluating the relation between early saliency and object saliency.

In any study of overt attention or object recognition, stimulus choice is critical. Stimulus category influences the prediction performance of saliency maps and other attention models (Einhäuser, Rutishauser, et al., 2006; Parkhurst et al., 2002; Peters et al., 2005; Privitera, Fujita, Chernyak, & Stark, 2005; Privitera & Stark, 2000). For our photographs of fairly complex everyday scenes, fixation prediction (58% AUC) is within the range of similar paradigms, which extends from 53% for the foliage images of Einhäuser, Kruse, Hoffmann, and König (2006) to the 68% of Peters et al. (2005). In terms of relating saliency maps to fixations, our images are typical.

The result that interesting objects are often accompanied by high saliency values was independently observed in a very recent analysis (Elazary & Itti, 2008) with a complementary approach: While we here use a well-controlled setting, these authors used a large database annotated by a huge set of—often unknown—observers (“LabelMe”). The fact that Elazary and Itti (2008) arrive at similar conclusions regarding the prediction of objects by saliency maps assures that this finding is not a consequence of our specific setting, tasks, or image material. Our data confirm the findings of Elazary and Itti on the relation between saliency and object naming in a controlled subject population and add the direct measurement of fixations. It should be noted, however, that neither our data nor Elazary and Itti's prove that there is a *causal* link between saliency and object recall. Saliency might merely be a correlate rather than a guiding principle of where objects are in natural scenes. The extent to which low-level features, such as those of the saliency model, guide object recall indeed causally will remain an interesting issue for further research.

Our prediction of object recall with object saliency suggests that models of attention may also model object properties in natural scenes. This opens several further lines of research. First, can attention models not only predict free recall, but also recognition performance under difficult conditions? Recent evidence suggests that a Bayesian model of surprising events, not only predicts attention allocation (Itti & Baldi, 2008), but also predicts human errors during natural scene recognition (Einhäuser, Mundhenk, Baldi, Koch, & Itti, 2007). Second, can we adapt low-level models of object recognition to predict attention allocation? Recently, Walther, Serre, Poggio, and Koch (2005) have proposed an architecture that shares features between attention and recognition. Third, can manipulating scene statistics dissociate attention and recognition? While these questions are beyond the scope of this paper, our data indicate that investigating the coupling of attention and recognition will be fruitful for understanding human vision under natural conditions and for modeling attention and recognition in real-world scenes.

Although frequently named objects are generally more fixated and more salient, the number of fixations on an object shows a larger variation for frequently named objects than for rarely named ones. In addition, if only one observer recalls a particular object, they have a slight tendency for a larger fraction of fixations on that object. Since we ask for “keywords,” we may have biased observers to name scene-diagnostic objects. It is therefore possible that rarely named objects could still be remembered, if they were specifically queried. In this view, less expected objects need more fixations (or more salience) to be named. This is in line with the idea that “surprising” (out of context/less expected) events draw attention, whether they deviate statistically (Itti & Baldi, 2008) or semantically (Friedman, 1979) from expectation. Indeed, “implausible” objects (i.e., objects that conflict with scene gist) tend to be recalled better (Pezdek, Whetstone, Reynolds, Askari, & Dougherty, 1989), although they are recognized worse (Davenport & Potter, 2004) and their effective field-of-view is smaller (DeGraef, 1998). Whether semantically implausible objects are fixated earlier or even “pop-out” (Loftus & Mackworth, 1978) has remained, however, controversial. Recent studies that use more complex scenes than Loftus and Mackworth (1978) and control the saliency of the critical item typically do not find an early preference to fixate implausible objects. Instead, they find that implausible objects are fixated longer, more likely to be fixated again (Henderson, Weeks, & Hollingworth, 1999), and are fixated earlier than plausible objects only after prolonged viewing (Underwood, Templeman, Lamming, & Foulsham, 2008) or if they appear while saccadic suppression suppresses bottom-up attention capture (Brockmole & Henderson, 2008). Under some experimental conditions, the recall of an item is improved by increased numbers of fixations on the object (Hollingworth & Henderson, 2002),

although this effect can be restricted to certain aspects of the item and depend on the methods of querying (Tatler, Gilchrist, et al., 2005). The effect that different object properties (saliency, object properties, fixation frequency, naming frequency, etc.) have on the ability of an observer to recall an item when queried will be an interesting issue for further investigation. The diversity of findings stresses that the querying for keywords in the present study and the unknown motivation of LabelMe participants in Elazary and Itti (2008) may yield substantially different results from other tasks, such as change detection or item recall.

In conclusion, we provide evidence that “interesting” objects, rather than early features, guide human attention. Some high-level scene interpretation is rapidly available to the visual system (Li et al., 2002; Rousselet et al., 2002; Thorpe, Fize, & Marlot, 1996), potentially faster or with

less effort than low-level concepts (Hochstein & Ahissar, 2002; Li et al., 2002). Together with the present data, this suggests another interesting speculation: Eye movements or spatial attention are by-products of object based attention or object recognition.

## Appendix A

### Task and fixation statistics

Here we address the effect of task on basic fixation statistics, duration, and location. During the 3-s image presentation in the “what” task, observers make on average  $10.0 \pm 0.4$  fixations (mean  $\pm$ SD across observers,

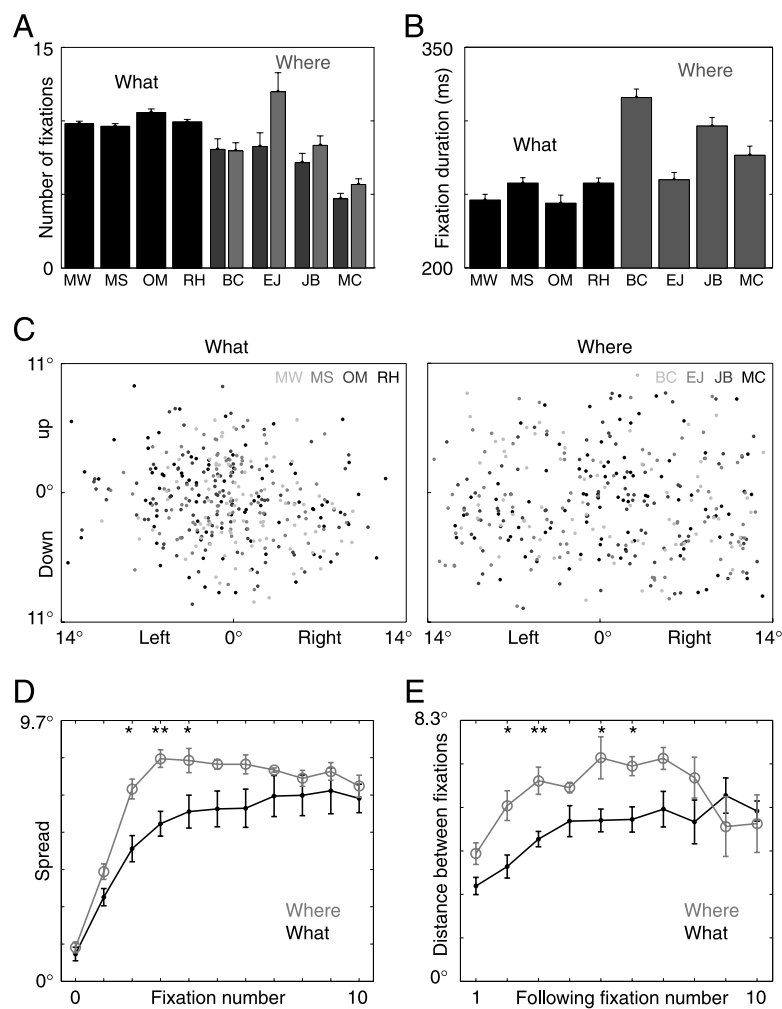


Figure A1. Basic fixation data. (A) Mean and SEM of number of fixations across images for individual observers. *Black bars*: observers “what” task (93 images per observer); *dark gray*: “where” task, target present (51 images); *light gray*: “where task,” target absent (42 images). (B) Mean and SEM fixation durations across images for each individual. (C) As an example of different fixation spread, all third fixations for each observer, rectangle corresponds to full image ( $1024 \times 768$  pixels); *left*: what-task; *right*: where-task. (D) Spread (root of sum of variances along cardinal axes) of fixations (each y tick mark corresponds to 50 pixels, i.e., about  $1.4^\circ$ ) on fixation number. Mean  $\pm$ SEM over observers in each task. 0 denotes initial fixation. (E) Distance between subsequent fixations, x-axis denotes fixation destination (e.g., “1” denotes distance between initial and first fixations).

all fixation counts exclude the initial, “0th” central fixation). The mean is smaller ( $7.7 \pm 2.0$ ) for the “where” task, in which observers terminate each trial themselves, but there is a larger variation: The standard deviation across images is  $1.9 \pm 0.3$  for “what,” but  $4.8 \pm 2.1$  for “where.” As expected, this high standard deviation arises from the fact that target-present trials have fewer fixations ( $7.0 \pm 1.6$ ) than target-absent trials ( $8.5 \pm 2.6$ ), and the inter-observer variation is substantial (Figure A1A).

Since trial duration differs between conditions (3 s versus self-termination), the relative number of fixations per unit time (or its inverse, the fixation duration) is of particular interest. In the “what” task, a fixation takes  $251 \pm 134$  ms (mean  $\pm$ SD across 3716 fixations). In the “where” task, a fixation takes  $286 \pm 153$  ms (2862 fixations), with no significant difference between target-present and target-absent trials ( $p = 0.24$ ,  $t$ -test; Figure A1B). The fixation duration difference in the “what” and “where” tasks is highly significant ( $p = 4 \times 10^{-23}$ ,  $t$ -test).

By experimental design, the spatial distribution of fixations shows a pronounced central bias (Figure A1C). Fixations in the “where” condition are, however, spread more widely than in the “what” condition. The standard deviation of each fixation’s location (square root of sum of variances of  $x$  coordinate and  $y$  coordinate) across images quantifies this spread. It is larger from the first to the tenth fixation in the “what” than in the “where” condition (Figure A1D). An alternative measure, the average distance between subsequent fixations, exhibits a similar time course (Figure A1E). In conclusion, duration and spatial distribution of fixations are task dependent.

## Appendix B

### Consistency of fixated locations

To investigate inter-observer consistency, we smooth the map of fixations with a one-degree wide Gaussian kernel to obtain a “fixation map.” We compute the map leaving out one observer and then predict that observer’s fixations with the map (Figure B1A). The map predicts fixations above chance (AUC greater 50%) in all images with the mean AUC over images ranging from  $82.9\% \pm 8.4\%$  (MW, mean  $\pm$ SD) to  $93.3\% \pm 5.4\%$  (MC) and a  $88.9\%$  mean across observers (Figure B1B, black). The random reassignment baseline yields a range of  $69.8\% \pm 11.0\%$  to  $79.8\% \pm 12.7\%$  (mean:  $75.7\%$ ; Figure B1B). This implies that a perfect model of average spatial distribution of fixations could predict up to  $75.7\%$  of fixations, *without knowledge of the actual stimulus*. Although this indicates that much across-observer consistency is caused by common spatial biases, the actual data exceed the random baseline significantly in all observers ( $p_{\max} = 4.8 \times 10^{-12}$ ,  $t$ -test). Consequently,

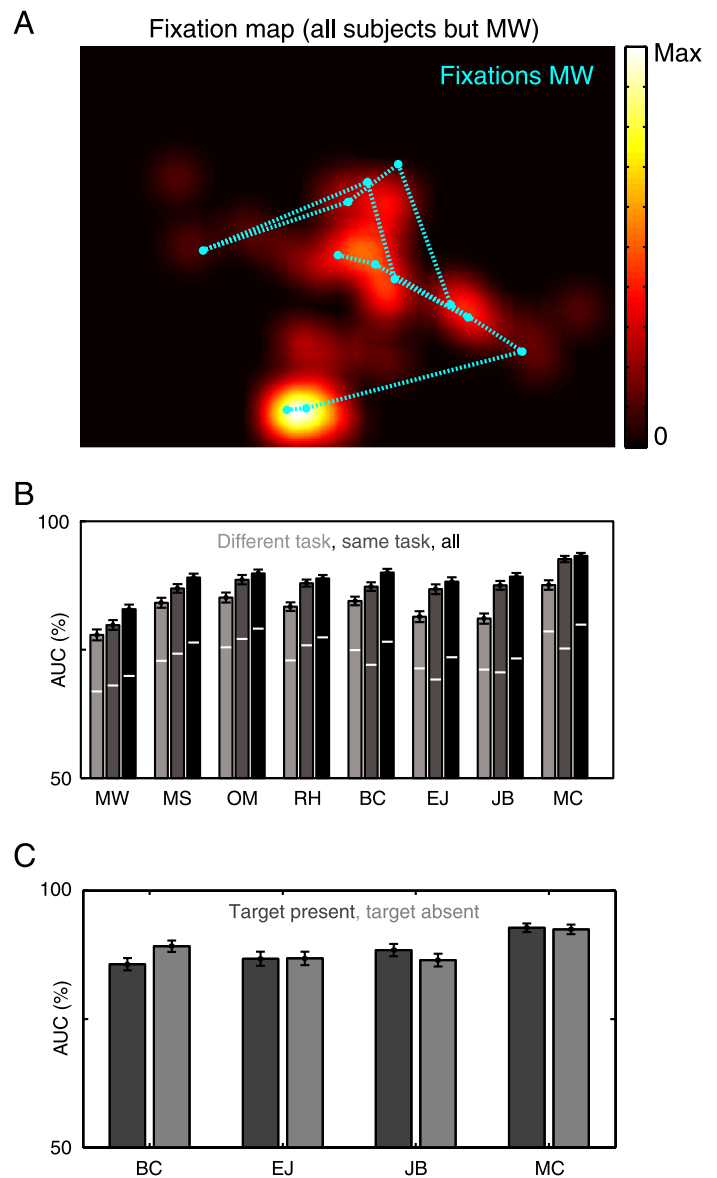


Figure B1. Inter-observer consistency. (A) “Fixation map” for image of Figure 2B with fixations of observer MW superimposed. Map is generated by filtering the fixations of 7 observers (all but MW) with a Gaussian of standard deviation  $1^\circ$ . (B) Area under ROC curve (AUC) for predicting fixations of one observer by the fixation map generated from other observers. *Left bar (light gray)*: fixation map generated from the 4 observers of the other task; *middle bar (dark gray)*: fixation map generated from the 3 other observers performing the same task; *right bar (black)*: fixation map generated from all 7 other observers. Mean and standard error across images for each observer. White lines denote results of random reassignment baseline. (C) Data for “where” observers of panel B separated by target-present (dark gray) and target-absent trials (light gray).

there is a large image specific (as compared to common bias) component to inter-observer consistency. Limiting the map calculation to within task slightly worsens predictions (Figure B1B, dark gray), on average by 1.7%



(what:  $1.9\% \pm 1.0\%$ ; where:  $1.6\% \pm 0.8\%$ ). This reduction is likely due to the smaller amount of data over which the map is computed, as the baseline shows a similar or larger drop (what:  $1.9\% \pm 0.3\%$ ; where:  $4.1\% \pm 0.9\%$ ). The significantly larger drop in the “where” condition ( $p = 0.004$ ,  $t$ -test), however, suggests that general spatial biases are slightly less relevant (as compared to image specific effects) in the “where” condition. This is in line with the faster spread of fixations during search (Figures A1D and A1E).

Predicting fixations with a map from the other task, is consistently worse than within task prediction (Figure B1B, light gray) and significantly worse (at  $p < 0.05$ ) in all but one observer (MW). This difference occurs even though the fixation map is based on 4 observers in the other task and only 3 in the same task. Nevertheless, even across tasks, the prediction is above chance for all but one image (JB for the image of an isolated chair, Figure 1, 4th item row 4). In the random reassignment baseline, there is no difference between prediction within and across tasks ( $p > 0.05$  for all observers), such that we have no evidence for a task modulation of the generic component (spatial bias) to inter-observer consistency. Within “where” observers, prediction does not consistently depend on target presence (Figure B1C), ruling out that fixations on or close to the target dominate inter-observer consistency in the “where” task. In summary, there is enough inter-observer consistency to predict another individual’s fixations, in spite of some task dependence.

## Appendix C

### Object recall statistics and inter-observer consistency

In each of the 93 images, there were between 6 and 16 objects recalled by at least one observer and  $10.5 \pm 2.3$  (mean  $\pm$ SD) on average (Table 1; Figure C1A). Across all 93 images, the 8 observers recalled 981 individual objects (objects are counted across images but once per image). Obvious synonyms were treated as the same object, while subcategories and parts were counted separately alongside the object. We denote the recall frequency of object  $j$  in image  $i$  by  $N(i, j)$ . Nearly half of the objects (457/981, 46.6%; Figure C1B) were recalled by only one individual (i.e.,  $N(i, j) = 1$ ), another 18.7% (183/981) only by two individuals ( $N(i, j) = 2$ ). Analyzing the 4 observers in each task separately, the objects recalled by a single observer account for more than half of the objects recalled (“what”: 338/590, 57.3%, Figure C1C; “where”: 418/794, 52.6%, Figure C1D). In each image, there was at least one object recalled by at least 4 observers ( $\max_j N(i, j) \geq 4$  for all  $i$ ; Figure C1E). In 64/93 (68.8%) images, there was an object, which at least 7 observers recalled ( $\max_j N(i, j) \geq 7$ ), and in 27/93 (29.0%) images, at least one object was recalled by

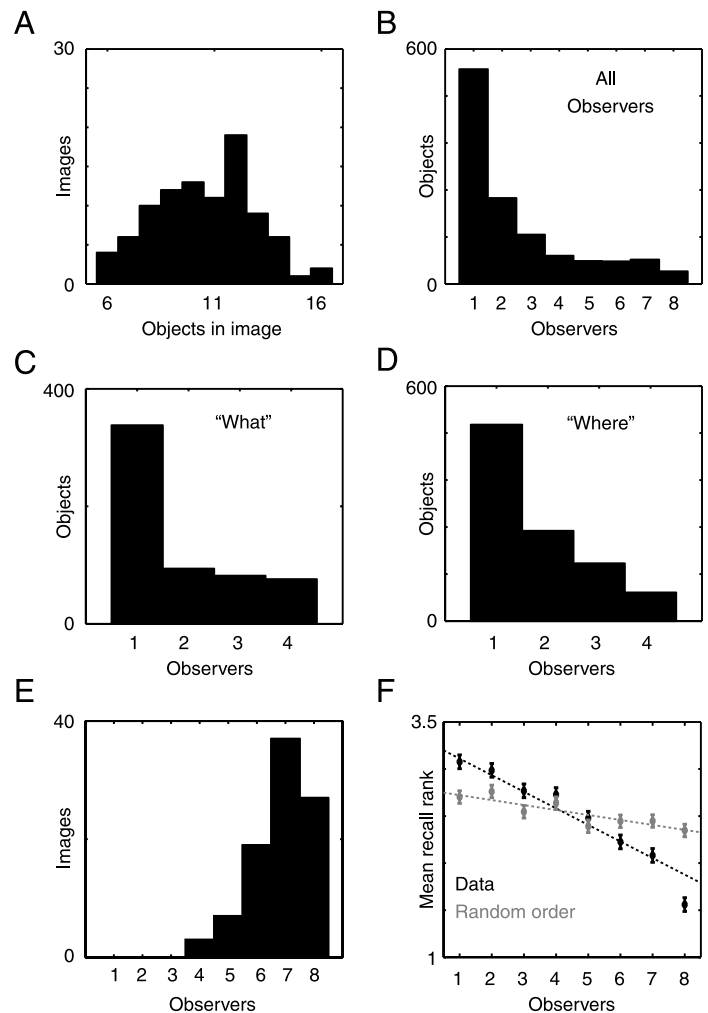


Figure C1. Object recall. (A) Number of different objects recalled in each image, histogram across images. (B) Incidences of naming frequencies, i.e., how many objects are named by 1, 2, ..., 8 observers. (C) As panel B for “what” task observers only. (D) As panel B for “where” task. (E). Recall frequency of most frequently recalled object in each image ( $\max_j N(i, j)$ ), histogram over images  $i$ . (F) Mean recall rank (1st, 2nd, ...) versus recall frequency. Black: Actual data, mean  $\pm$ SEM across objects. Note that correlation values in the text treat each of the 981 objects as individual data point, which is also used for the fits in the figures. Correlating naming frequency to mean values (i.e., correlating 8 data points) would result in much high correlation values, here:  $r = -0.97$  ( $p = 6 \times 10^{-5}$ ). Gray: random baseline: For each image, the order of all objects a given observer recalled is randomly shuffled while preserving the total number of object he/she recalled in this image. For extreme values of recall frequency (1, 2, 3, 7, 8), real data are significantly different from the baseline ( $p = 2 \times 10^{-5}$ ;  $p = 0.03$ ;  $p = 0.048$ ;  $p = 0.002$ ;  $p = 2 \times 10^{-8}$ ).

all 8 observers ( $\max_j N(i, j) = 8$ ; Figure C1E). This means that in most images, there is at least one “characteristic object,” an object that is recalled by most of the observers. This motivates the search for distinctive properties of these characteristic objects.

The order in which a given object is recalled presents an alternative measure as to how characteristic or important an object is for a scene. There is a highly significant correlation between recall frequency and recall order ( $r = -0.31$ ,  $p = 2 \times 10^{-23}$ ; Figure C1F, black). Actual recall ranks differ significantly from a baseline that corrects for having no lower limit on the number of objects an individual recalls (Figure C1F, gray). That is, individuals name frequently recalled objects earlier than idiosyncratic objects. This motivates to restrict analysis on recall frequency.

## Appendix D

### Recall prediction by combination of object properties

Total object saliency combines the saliency of an object and its area to a common measure. Consequently, both measures are tightly correlated ( $r = 0.76$ ,  $p = 3 \times 10^{-186}$ ); similarly, boundary length and area are trivially coupled, with larger area implying a longer boundary ( $r = 0.63$ ,  $p = 8 \times 10^{-110}$ ). As only parts of objects that fall within the image boundary are used to determine its center of mass, large objects have a bias toward the center, reflected in a correlation between center distance and area ( $r = -0.30$ ,  $p = 4 \times 10^{-22}$ ). Maximum object saliency is correlated to all these measures, trivially to total object saliency ( $r = 0.56$ ;  $p = 1.5 \times 10^{-81}$ ) and also to object area ( $r = 0.39$ ,  $p = 3.4 \times 10^{-36}$ ). The latter correlation can partly be understood as a consequence of the sparsity of saliency maps: Peaks are rare, while low values occur frequently; hence, larger objects have a slightly better chance to “capture” the peak.

How well does a linear combination of these properties predict recall frequency? Combining area and total object saliency by performing discrimination along the first

principal axis of all data yields slightly better results than either measure alone: AUCs range from 69.2% (named once versus named twice or more; Figure D1) to 86.2% (once versus 8 times). Similar unsupervised inclusion of the other measures or combining more than 2 measures does not yield better prediction performance (Figure D1).

## Acknowledgments

This material is based upon work supported under a National Science Foundation Graduate Research Fellowship, National Institute of Mental Health grant T32MH019138, Office of Naval Research grant N00014-06-1-0734, National Institutes of Health grant R01 DA022777, and Swiss National Science Foundation fellowship PA00A-111447.

Commercial relationships: none.

Corresponding author: Wolfgang Einhäuser.

Email: wet@physik.uni-marburg.de.

Address: Department of Neurophysics, Philipps-University Marburg, Renthof 7, 35032, Marburg, Germany.

## Footnotes

<sup>1</sup>We here use “recognition” in its broadest possible meaning, which may include detection, the actual recognition process, inscription, and consolidation into visual short-term memory (VSTM), and the retrieval from VSTM. The different phenomena listed impair different parts of this processing chain and we expect that a lot of the seeming conflict in the literature is resolved by a more precise distinction.

<sup>2</sup>Owing to the extension of objects, which is large compared to the resolution of the saliency map, and to the

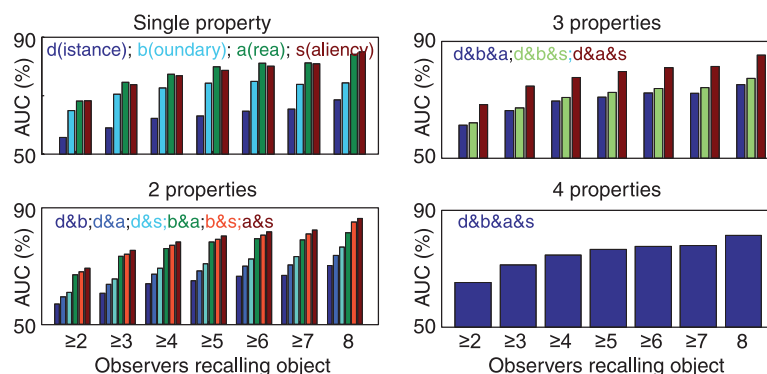


Figure D1. Recall prediction by various object properties and their linear combination Area under curve for the prediction of recall using all combinations of four image properties: distance from center, outline (boundary) length, fractional area, total object saliency. Colors of bars denote measures as given in panel legend. Measures including saliency and/or area outperform the other measures.

limited resolution of the eye tracking device (see below), analyzing the data at the reduced resolution does not affect any of the reported results qualitatively.

<sup>3</sup>We assume the same objects being named, but the most salient object selected at random. If there are  $N_i$  objects in image  $i$ , of which  $n_i$  are named most frequently, the probability of picking one of these objects is  $p_i = n_i / N_i$ . The probability that for exactly  $k$  images the randomly picked object is among the most frequently named is  $P_L(X = k) = \sum_{V \in \Omega_k^L} \prod_{i \in V} p_i \prod_{i \in \bar{V}} (1 - p_i)$ , where  $\Omega_k^L$  denotes the set of all  $k$ -element subsets of  $\{1, \dots, L\}$ ,  $V$  is one particular subset,  $\bar{V}$  is the complement of  $V$  with respect to  $\{1, \dots, L\}$ , and  $L$  is the number of images. The  $p$  values in the text correspond to  $P_{93}(X \geq 34)$  for all images and  $P_{77}(X \geq 28)$  for the subset with  $n_i = 1$ . With  $\binom{93}{34} > 10^{25}$  and  $\binom{77}{28} > 10^{20}$  this is infeasible to compute analytically, therefore we performed the simulation.

<sup>4</sup>Since the saliency map has a lower resolution than the image, we consider the peak to be at the center of the  $16 \times 16$  pixel region of the image, for which the saliency map takes its maximum value (1). In two maps, saliency map takes the maximum value at two locations, in the case of connectedness (1 image), we place the peak between those two pixels, in the unconnected case, we consider both peaks and correct the baseline probability accordingly  $p' = p + p(1 - p) = 2p - p^2$ .

## References

- Armstrong, K. M., Fitzgerald, J. K., & Moore, T. (2006). Changes in visual receptive fields with microstimulation of frontal cortex. *Neuron*, *50*, 791–798. [PubMed] [Article]
- Bichot, N. P., Rossi, A. F., & Desimone, R. (2005). Parallel and serial neural mechanisms for visual search in macaque area V4. *Science*, *308*, 529–534. [PubMed]
- Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*, 433–436. [PubMed]
- Brockmole, J. R., & Henderson, J. M. (2008). Prioritizing new objects for eye fixation in real-world scenes: Effects of object-scene consistency. *Visual Cognition*, *16*, 375–390.
- Buswell, G. T. (1935). *How people look at pictures. A study of the psychology of perception in art*. Chicago: University of Chicago Press.
- Carmi, R., & Itti, L. (2006). Visual causes versus correlates of attentional selection in dynamic scenes. *Vision Research*, *46*, 4333–4345. [PubMed]
- Cerf, M., Harel, J., Einhäuser, W., & Koch, C. (2008). Predicting human gaze using low-level saliency combined with face detection. *Advances in Neural Information Processing Systems*, *20*, 241–248.
- Cornelissen, F. W., Peters, E. M., & Palmer, J. (2002). The Eyelink Toolbox: Eye tracking with MATLAB and the Psychophysics Toolbox. *Behavior Research Methods, Instruments, & Computers*, *34*, 613–617. [PubMed]
- Davenport, J. L., & Potter, M. C. (2004). Scene consistency in object and background perception. *Psychological Science*, *15*, 559–564. [PubMed]
- De Graef, P. (1998). Prefixational object perception in scenes: Objects popping out of schemas. In G. Underwood (Ed.), *Eye guidance in reading and scene perception* (pp. 315–338). Oxford: Elsevier.
- De Graef, P. (2005). Semantic effects on object selection in real-world scene perception. In G. Underwood (Ed.), *Cognitive processes in eye guidance* (pp. 189–212). Oxford: Oxford University Press.
- Dickinson, S., Christensen, H., Tsotsos, J., & Olofsson, G. (1997). Active object recognition integrating attention and viewpoint control. *Computer Vision and Image Understanding*, *63*, 239–260.
- Einhäuser, W., Hipp, J., Eggert, J., Körner, E., & König, P. (2005). Learning viewpoint invariant object representations using a temporal coherence principle. *Biological Cybernetics*, *93*, 79–90. [PubMed]
- Einhäuser, W., Koch, C., & Makeig, S. (2007). The duration of the attentional blink in natural scenes depends on stimulus category. *Vision Research*, *47*, 597–607. [PubMed] [Article]
- Einhäuser, W., & König, P. (2003). Does luminance-contrast contribute to a saliency map for overt visual attention? *European Journal of Neuroscience*, *17*, 1089–1097. [PubMed]
- Einhäuser, W., Kruse, W., Hoffmann, K. P., & König, P. (2006). Differences of monkey and human overt attention under natural conditions. *Vision Research*, *46*, 1194–1209. [PubMed]
- Einhäuser, W., Mundhenk, T. N., Baldi, P., Koch, C., & Itti, L. (2007). A bottom-up model of spatial attention predicts human error patterns in rapid scene recognition. *Journal of Vision*, *7*(10):6, 1–13, <http://journalofvision.org/7/10/6/>, doi:10.1167/7.10.6. [PubMed] [Article]
- Einhäuser, W., Rutishauser, U., Frady, E. P., Nadler, S., König, P., & Koch, C. (2006). The relation of phase noise and luminance contrast to overt attention in complex visual stimuli. *Journal of Vision*, *6*(11):1, 1148–1158, <http://journalofvision.org/6/11/1/>, doi:10.1167/6.11.1. [PubMed] [Article]
- Einhäuser, W., Rutishauser, U., & Koch, C. (2008). Task-demands can immediately reverse the effects of sensory-driven saliency in complex visual stimuli. *Journal of Vision*, *8*(2):2, 1–19, <http://journalofvision.org/8/2/2/>, doi:10.1167/8.2.2. [PubMed] [Article]



- Elazary, L., & Itti, L. (2008). Interesting objects are visually salient. *Journal of Vision*, 8(3):3, 1–15, <http://journalofvision.org/8/3/3/>, doi:10.1167/8.3.3. [PubMed] [Article]
- Evans, K. K., & Treisman, A. (2005). Perception of objects in natural scenes: Is it really attention free? *Journal of Experimental Psychology: Human Perception and Performance*, 31, 1476–1492. [PubMed]
- Fei-Fei, L., Fergus, R., & Perona, P. (2004). Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories. *CVPR 2004, Workshop on Generative-Model Based Vision*.
- Fergus, R., Perona, P., & Zisserman, A. (2003). Object class recognition by unsupervised scale-invariant learning. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2, 246–271.
- Friedman, A. (1979). Framing pictures: The role of knowledge in automatized encoding and memory for gist. *Journal of Experimental Psychology: General*, 108, 316–355. [PubMed]
- Henderson, J. M., Brockmole, J. R., Castelano, M. S., & Mack, M. (2006). Visual saliency does not account for eye-movements during visual search in real-world scenes. In R. van Gompel, M. Fischer, W. Murray, & R. Hill (Eds.), *Eye movement research: Insights into mind and brain* (pp. 537–562). Oxford: Elsevier.
- Henderson, J. M., Weeks, Jr., P. A., & Hollingworth, A. (1999). The effects of semantic consistency on eye movements during complex scene viewing. *Journal of Experimental Psychology: Human Perception and Performance*, 25, 210–228.
- Henderson, J. M., Williams, C. C., Castelano, M. S., & Falk, R. J. (2003). Eye movements and picture processing during recognition. *Perception & Psychophysics*, 65, 725–734. [PubMed]
- Hershler, O., & Hochstein, S. (2005). At first sight: A high-level pop out effect for faces. *Vision Research*, 45, 1707–1724. [PubMed]
- Hershler, O., & Hochstein, S. (2006). With a careful look: Still no low-level confound to face pop-out. *Vision Research*, 46, 3028–3035.
- Hochstein, S., & Ahissar, M. (2002). View from the top: Hierarchies and reverse hierarchies in the visual system. *Neuron*, 36, 791–804. [PubMed] [Article]
- Hollingworth, A., & Henderson, J. M. (2002). Accurate visual memory for previously attended objects in natural scenes. *Journal of Experimental Psychology: Human Perception and Performance*, 28, 113–136.
- Itti, L., & Baldi, P. (2008). Bayesian surprise attracts human attention. *Vision Research*. [PubMed]
- Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, 40, 1489–1506. [PubMed]
- James, W. (1890). *Principles of psychology*. New York: Holt.
- Kanwisher, N. G. (1987). Repetition blindness: Type recognition without token individuation. *Cognition*, 27, 117–143. [PubMed]
- Kayser, C., Nielsen, K. J., & Logothetis, N. K. (2006). Fixations in natural scenes: Interaction of image structure and image content. *Vision Research*, 46, 2535–2545. [PubMed]
- Koch, C., & Ullman, S. (1985). Shifts in selective visual attention: Towards the underlying neural circuitry. *Human Neurobiology*, 4, 219–227. [PubMed]
- Krieger, G., Rentschler, I., Hauske, G., Schill, K., & Zetsche, C. (2000). Object and scene analysis by saccadic eye-movements: An investigation with higher-order statistics. *Spatial Vision*, 13, 201–214. [PubMed]
- Land, M. F., & Hayhoe, M. (2001). In what ways do eye movements contribute to everyday activities? *Vision Research*, 41, 3559–3565. [PubMed]
- LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86, 2278–2324.
- Li, F. F., VanRullen, R., Koch, C., & Perona, P. (2002). Rapid natural scene categorization in the near absence of attention. *Proceedings of the National Academy of Sciences of the United States of America*, 99, 9596–9601. [PubMed] [Article]
- Li, Z. (2002). A saliency map in primary visual cortex. *Trends in Cognitive Sciences*, 6, 9–16. [PubMed]
- Loftus, G. R., & Mackworth, N. H. (1978). Cognitive determinants of fixation location during picture viewing. *Journal of Experimental Psychology: Human Perception and Performance*, 4, 565–572. [PubMed]
- Mannan, S. K., Ruddock, K. H., & Wooding, D. S. (1996). The relationship between the locations of spatial features and those of fixations made during visual examination of briefly presented images. *Spatial Vision*, 10, 165–188. [PubMed]
- Mannan, S. K., Ruddock, K. H., & Wooding, D. S. (1997). Fixation patterns made during brief examination of two-dimensional images. *Perception*, 26, 1059–1072. [PubMed]
- Mazer, J. A., & Gallant, J. L. (2003). Goal-related activity in V4 during free viewing visual search. Evidence for a ventral stream visual salience map. *Neuron*, 40, 1241–1250. [PubMed] [Article]
- Mel, B. W. (1997). SEEMORE: Combining color, shape, and texture histogramming in a neurally inspired

- approach to visual object recognition. *Neural Computation*, 9, 777–804. [PubMed]
- Nakayama, K. (1990). The iconic bottleneck and the tenuous link between early visual processing and perception. In C. Blakemore (Ed.), *Vision: Coding and efficiency* (pp. 135–149). Cambridge: Cambridge University Press.
- Navalpakkam, V., & Itti, L. (2007). Search goal tunes visual features optimally. *Neuron*, 53, 605–617. [PubMed] [Article]
- Neisser, U., & Becklen, L. (1975). Selective looking: Attending to visually specified events. *Cognitive Psychology*, 7, 480–494.
- Nelson, W. W., & Loftus, G. R. (1980). The functional visual field during picture viewing. *Journal of Experimental Psychology: Human Learning and Memory*, 6, 391–399. [PubMed]
- Ogawa, T., & Komatsu, H. (2006). Neuronal dynamics of bottom-up and top-down processes in area V4 of macaque monkeys performing a visual search. *Experimental Brain Research*, 173, 1–13. [PubMed]
- Parker, R. E. (1978). Picture processing during recognition. *Journal of Experimental Psychology: Human Perception and Performance*, 4, 284–293. [PubMed]
- Parkhurst, D., Law, K., & Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision Research*, 42, 107–123. [PubMed]
- Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, 10, 437–442. [PubMed]
- Peters, R. J., Iyer, A., Itti, L., & Koch, C. (2005). Components of bottom-up gaze allocation in natural images. *Vision Research*, 45, 2397–2416. [PubMed]
- Pezdek, K., Whetstone, T., Reynolds, K., Ashkari, N., & Dougherty, T. (1989). Memory for real-world scenes: The role of consistency with schema expectation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15, 587–595.
- Pomplun, M. (2006). Saccadic selectivity in complex visual search displays. *Vision Research*, 46, 1886–1900. [PubMed]
- Privitera, C. M., Fujita, T., Chernyak, D., & Stark, L. W. (2005). On the discriminability of hROIs, human visually selected regions-of-interest. *Biological Cybernetics*, 93, 141–152. [PubMed]
- Privitera, C. M., & Stark, L. W. (2000). Algorithms for defining visual regions-of-interest: Comparison with eye fixations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22, 970–982.
- Rao, R. P., Zelinsky, G. J., Hayhoe, M. M., & Ballard, D. H. (2002). Eye movements in iconic visual search. *Vision Research*, 42, 1447–1463. [PubMed]
- Raymond, J. E., Shapiro, K. L., & Arnell, K. M. (1992). Temporary suppression of visual processing in an RSVP task: An attentional blink? *Journal of Experimental Psychology: Human Perception and Performance*, 18, 849–860. [PubMed]
- Reinagel, P., & Zador, A. M. (1999). Natural scene statistics at the centre of gaze. *Network*, 10, 341–350. [PubMed]
- Rensink, R. A., O'Regan, J. K., & Clark, J. J. (1997). To see or not to see: The need for attention to perceive changes in scenes. *Psychological Science*, 8, 368–373.
- Riesenhuber, M., & Poggio, T. (2002). Neural mechanisms of object recognition. *Current Opinion in Neurobiology*, 12, 162–168. [PubMed]
- Rizzolatti, G., Riggio, L., Dascola, I., & Umiltà, C. (1987). Reorienting attention across the horizontal and vertical meridians: Evidence in favor of a premotor theory of attention. *Neuropsychologia*, 25, 31–40. [PubMed]
- Rousselet, G. A., Fabre-Thorpe, M., & Thorpe, S. J. (2002). Parallel processing in high-level categorization of natural images. *Nature Neuroscience*, 5, 629–630. [PubMed]
- Rutishauser, U., Walther, D., Koch, C., & Perona, P. (2004). Is bottom-up attention useful for object recognition? *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2, 37–44.
- Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., & Poggio, T. (2007). Robust object recognition with cortex-like mechanisms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29, 411–426. [PubMed]
- Shore, S., Tillman, L., & Schmidt-Wulffen, S. (2004). *Stephen shore: Uncommon places: The Complete works*. New York: Aperture.
- Simons, D. J. (2000). Attentional capture and inattention blindness. *Trends in Cognitive Sciences*, 4, 147–155. [PubMed]
- Tatler, B. W. (2007). The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision*, 7(14):4, 1–17, <http://journalofvision.org/7/14/4/>, doi:10.1167/7.14.4. [PubMed] [Article]
- Tatler, B. W., Baddeley, R. J., & Gilchrist, I. D. (2005). Visual correlates of fixation selection: Effects of scale and time. *Vision Research*, 45, 643–659. [PubMed]
- Tatler, B. W., Gilchrist, I. D., & Land, M. F. (2005). Visual memory for objects in natural scenes: From fixations to object files. *Quarterly Journal of Experimental Psychology A: Human Experimental Psychology*, 58, 931–960. [PubMed]

- Tatler, B. W., Wade, N. J., & Kaulard, K. (2007). Examining art: Dissociating pattern and perceptual influences on oculomotor behaviour. *Spatial Vision*, 21, 165–184. [PubMed]
- Thompson, K. G., & Bichot, N. P. (2005). A visual salience map in the primate frontal eye field. *Progress in Brain Research*, 147, 251–262. [PubMed]
- Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature*, 381, 520–522. [PubMed]
- Torralba, A., Oliva, A., Castelhano, M. S., & Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review*, 113, 766–786. [PubMed]
- Tsotsos, J. K., Culhane, S., Wai, W., Lai, Y., Davis, N., & Nuflo, F. (1995). Modeling visual attention via selective tuning. *Artificial Intelligence*, 78, 507–547.
- Underwood, G., Foulsham, T., van Loon, E., Humphreys, L., & Bloyce, J. (2006). Eye movements during scene inspection: A test of the saliency map hypothesis. *European Journal of Cognitive Psychology*, 18, 321–343.
- Underwood, G., Templeman, E., Lamming, L., & Foulsham, T. (2008). Is attention necessary for object identification? Evidence from eye movements during the inspection of real-world scenes. *Consciousness and Cognition*, 17, 159–170. [PubMed]
- VanRullen, R. (2006). On second glance: Still no high-level pop-out effect for faces. *Vision Research*, 46, 3017–3027. [PubMed]
- Wallis, G., & Rolls, E. T. (1997). Invariant face and object recognition in the visual system. *Progress in Neurobiology*, 51, 167–194. [PubMed]
- Walther, D., Serre, T., Poggio, T., & Koch, C. (2005). Modeling feature sharing between object detection and top-down attention [Abstract]. *Journal of Vision*, 5(8):1041, 1041a, <http://journalofvision.org/5/8/1041/>, doi:10.1167/5.8.1041.
- Yarbus, A. L. (1967). *Eye movements and vision*. New York: Plenum.